Статистика речи и автоматический анализ текста

1980

АКАДЕМИЯ НАУК СССР
ИНСТИТУТ ЯЗЫКОЗНАНИЯ

СТАТИСТИКА РЕЧИ И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТА

ЛЕНИНГРАД
«НАУКА»
ЛЕНИНГРАДСКОЕ ОТДЕЛЕНИЕ
1980

С $\frac{70101\text{-}566}{042\ (02)\text{-}80}$ 507.80.1503000000.

ПРЕДИСЛОВИЕ

Предлагаемый вниманию сборник представляет собой шестой выпуск заглавной серии исследований, публикуемых общесоюзной группой «Статистика речи» (см.: СтР, 1968; СтРААТ, 1971; 1973; 1974; 1979). Все эти исследования подчинены общей идее построения систем автоматического анализа и синтеза текста, написанного на естественном языке — языке, представляющем собой нечеткую систему, которая охватывает размытые множества, состоящие из нечетких лингвистических объектов.

Эта концепция разрабатывается и в статьях настоящего сборника. В первом разделе, в статье Р. Г. Пиотровского, Н. В. Аникиной, Т. А. Аполлонской и других описывается построение многоуровневой системы формального распознавания смысла текстовых единиц и устранения их многозначности, а также экспериментальная реализация этой системы на ЕС ЭВМ. Особенность этой системы состоит в том, что при ее построении учтен переход от нечеткости естественного языка к классическим множествам и объектам искусственного языка ЭВМ.

Во второй части «Статистика речи» собраны работы, трактующие вопросы такой организации исходной выборки, которая могла бы обеспечить объективность лингвостатистического эксперимента, а также выдачу надежных и достаточно типизированных статистических данных о построении текста (см. статьи С. А. Шубика, Н. А. Козинцевой и А. Г. Козинцева, А. В. Грошевой). Полученная с помощью предлагаемых в этих статьях приемов статистика может служить средством формального описания нечеткости естественного языка в машине.

Третья часть «Автоматический анализ текста» посвящена, с одной стороны, формально-лингвистическим приемам снятия

неоднозначности (омонимии) языковых объектов и представления их в удобном для машины виде (см. статью Л. В. Малаховского), а с другой — вопросам хранения и переработки в современных ЭВМ таких объемов лингвистической и энциклопедической информации, которые необходимы для осуществления не только простых процедур АПТ типа пословно-пооборотного перевода, но также и таких сложных операций, как распознавание смысла, семантико-синтаксический анализ («синтез»), ситуативное реферирование текста, человеко-машинный диалог и т. п. (см. статью Е. М. Лукьяновой).

Часть I

ФОРМАЛЬНОЕ РАСПОЗНАВАНИЕ СМЫСЛА

Р. Г. Пиотровский, Н. В. Аникина, Т. А. Аполлонская, В. Н. Билан, М. Н. Боркун, М. М. Лесохин, Е. А. Шингарева

ФОРМАЛЬНОЕ РАСПОЗНАВАНИЕ СМЫСЛА ТЕКСТА

§ 1. Введение. Общие положения теории формального распознавания смысла

Одной из центральных проблем современной инженерной лингвистики и лингвистических аспектов искусственного разума (ИР) является развитие теории формального распознавания смысла (ФРС) текста и разработка реализующего ее лингвоматематического и алгоритмического аппарата. Поскольку ФРС оперирует с лингвистическими объектами, имеющими семиотическую природу, и опирается на формальную языковую модель, в ней различаются четыре семиотических уровня: уровень денотата, уровень концепта, уровень означающего (имени) и прагматический уровень. Все базовые понятия ФРС, к определению которых мы сейчас переходим, будут рассматриваться с учетом указанных уровней.

1. Исходным понятием при ФРС любого лингвистического объекта, имеющего знаковую природу (морфема, словоформа (с/ф), слово, словосочетание (с/с), предложение, целостный текст), является понятие предметной области (ПО). ПО, будучи отражением в сознании отдельного человека или коллектива определенного участка объективной действительности, представляет денотативный уровень модели ФРС. Каких-либо ограничений при выборе ПО не существует: этот выбор производится соответственно задачам данного лингвистического или инженерно-лингвистического исследования. Предметной областью может стать отрасль знаний (например, специальность «современные методы окраски в машиностроении» или «опухоли человека» — Арзикулов и др., 1978), совокупность некоторых понятий, выраженных с помощью лексической единицы (ЛЕ) (например, предметная область «сутки», включающая понятия

«утро», «день», «вечер», «ночь»), или система локально-темпоральных отношений, воплощаемая обычно предложно-падежными средствами языка (Горбунов, 1978).

ПО делится в сознании отдельного человека или коллектива на смысловые «районы», «подрайоны» и т. д. На нижнем уровне иерархии располагаются элементарные денотаты (объекты, ситуации, отношения). Следует всегда помнить, что деление ПО может происходить по-разному у разных людей и коллективов. Если в условиях передачи сообщения, т. е. в ходе информационного процесса (ИП), передатчик и приемник используют разное разбиение ПО, то это неизбежно приводит к потерям информации, содержащейся в сообщении.

2. Поскольку каждая ПО достаточна сложна, охватывает слишком большое количество объектов и таким образом недоступна прямому наблюдению и копированию в памяти лингвистического автомата, то для осуществления ФРС приходится строить аналог (модель), более или менее наглядно и конструктивно отражающий основные свойства этой предметной области. Такую модель, развертываемую на уровне концептов, мы будем называть семантическим пространством (СП). Каждому из фрагментов ПО в ее модели задан соответствующий аналог. Так, например, районам ПО сопоставляются области СП, подрайонам соотнесены подобласти, а элементарным денотатам — точки СП или дуги, соединяющие эти точки.

3. В результате формируется уровневая иерархия СП, где верхние уровни отводятся для областей и подобластей, а на нижних располагаются точки СП. Каждый уровень СП описывается набором признаков. Чем выше этот уровень (микрообласть — подобласть — область), тем более общий характер носят признаки. И, наоборот, при движении от верхних уровней к нижним признаки получают более индивидуальный характер. При этом признаки более низкого уровня наследуют признаки предыдущих уровней иерархии. В результате внутри семантического пространства формируется поле признаков (ПП).

4. Поле признаков является рабочим инструментом при описании уровня концептов. При построении СП семантические признаки (СемП) выступают в роли организующих элементов, вокруг которых группируются области, подобласти, микрообласти СП. Что же касается конкретных концептов (точек СП), то признаки обычно выступают здесь в качестве наиболее актуальной, релевантной с точки зрения получателя (или отправителя) и ситуации («фокальной») семы для каждого из этих концептов (Shelling, 1960; Тоом, 1976; Щедровицкий, 1974; Шингарева, 1979).

5. Для реального осуществления ФРС необходимо соотнести построенную нами концептуальную модель (т. е. СП вместе с ПП) с планом выражения языка. Эта задача решается путем соотнесения областей, подобластей, микрообластей, . . ., точек СП с лексикой и фразеологией конкретного языка. В ходе этого соотнесения вырабатываются эталоны, представляющие собой формальное описание фрагментов СП, построенное на основе его признаков.

6. Каждая область (подобласть, . . ., точка) СП должна иметь также свое обозначение (метку), в качестве которой могут выступать как символы искусственного языка (ИЯ), так и лингвистические единицы (слова, словосочетания, предложения). Полный перечень этих меток мы будем называть алфавитом меток.

7. Алфавит представляет собой простой перечень меток областей, подобластей, микрообластей и точек СП. Этот список является вполне удовлетворительным средством для первичного описания СП. Если же возникает необходимость более глубокого структурного представления СП, то производится упорядочение меток друг относительно друга, в результате чего возникает семантическая сеть (СС), составляющая ту упрощенную концептуальную модель СП, которая вводится в лингвистический автомат (ЛА) для осуществления ФРС.

8. Чаще всего СС реализуется в виде древесного графа (тезауруса), в вершине которого помещается название ПО и соответственно СП, а в остальных узлах размещены метки областей, подобластей и т. д., выраженные словами и словосочетаниями. Эти метки воплощают понятия ПО. Каждый узел тезауруса отмечается специальным кодом, структура которого указывает на место узла в дереве. Узлы графа соединяются дугами, которые показывают на отношения между понятиями ПО. Эта использованная в настоящей работе организация позволяет применить к описанию СП язык алгебры отношений. Последнее свидетельствует о возможности строгого математического представления не только лингвистического описания ПО (см. § 3), но и самого алгоритма ФРС.

9. Смоделированная с помощью тезаурусного графа совокупность признаков, областей (подобластей, микрообластей, . . ., точек) СП, их эталонов и меток, размещаемых в базе данных ЛА, составляет вместе с алгоритмом их применения информационный язык системы ФРС.

10. Формальное распознавание смысла лингвистических единиц состоит в том, что распознающая система (лингвистический автомат или человек), пользуясь алгоритмической процедурой, относит поступивший на вход системы лингвистический объект к одному или нескольким заложенным в ее памяти областям (подобластям, . . ., точкам), выдавая при этом метку данной области (подобласти, . . ., точки) СП. Выданная системой метка рассматривается *как распознанный смысл той лингвистической* единицы, которая подвергалась формальному распознаванию.

§ 2. МП и проблема многозначности

Как уже говорилось, все формы автоматической переработки текста (АПТ), предусматривающие семантические операции, в том числе и МП, опираются на формальное распознавание смысла. Широко реализуемый в настоящее время пословный и необоротный МП представляет наиболее простую форму ФРС. Здесь каждая лексическая единица (т. е. слово, словоформа или словосочетание), заложенная в автоматическом словаре (АС), представляет собой аналог точки СП. Таким образом, уровень концептов в АС представляется в простейшем случае в виде множества точек СП, каждая из которых в плане выражения соответствует ЛЕ входного языка. В ходе распознавания смысла входного текста ФРС сопоставляет каждой входной ЛЕ (точке СП) ее метку, в роли которой выступает *выходной (в нашем случае — русский) эквивалент*. Весь процесс формального распознавания сводится здесь к отождествлению текстового словоупотребления или словосочетания с соответствующей входной ЛЕ в автоматическом словаре, а затем к выдаче метки, включающей один или несколько русских эквивалентов (Пиотровский, 1979, с. 77—79). Такая упрощенная модель концептуального уровня ПО, в которой СП представлено лишь в виде множества точек, каждой из которых сопоставляется ее метка на выходном языке, не может выдавать семантически адекватного перевода. Она позволяет получить весьма приблизительное распознавание смысла входного текста.

Гораздо более сложной и интересной с методологической и инженерно-лингвистической точек зрения задачей является двуязычное распознавание смысла многозначных лексических единиц, опирающееся на анализ их контекстного окружения (Dreyfus, 1972/1978, с. 70—72, 171—189; Пиотровский, 1975, с. 83—91, 278—286) и сочетающееся с использованием тезаурусного моделирования СП. Решению именно этой задачи и будет посвящена настоящая статья. Построение каждой системы ФРС предусматривает решение трех подзадач.

Первая подзадача заключается в выделении лингвистических объектов и признаков, относящихся к данной предметной области. Эти объекты и признаки, имеющие нечеткую природу, должны быть предусмотрены в системе ФРС как четкие семиотические объекты.

Вторая подзадача состоит в создании модели ПО, т. е. в развертывании СП, его областей, подобластей, . . ., точек, в формировании поля признаков, задания эталонов, алфавита меток и *построении тезауруса*.

Третья подзадача заключается в построении алгоритма распознавания смысла интересующих нас текстовых единиц, программировании этого алгоритма и его реализации на ЭВМ.

Первые две задачи выполняются на доалгоритмическом этапе исследования. Последняя задача образует его алгоритмический этап.

§ 3. Математическая модель формального распознавания смысла многозначных ЛЕ в рамках машинного перевода

3.1. Сгущение нечетких лингвистических множеств

Все десигнаты таких лингвистических объектов, как слова, словосочетания, представляя собой совокупность конкретных денотатов, являются нечеткими множествами, имеющими размытые границы. Это значит, что в отличие от классических «четких» множеств, где каждый элемент либо принадлежит множеству ($a \in A'$), либо не принадлежит ему ($a \notin A'$), в нечетких множествах принадлежность элемента к множеству оценивается некоторым коэффициентом μ, принимающим значения от 0 до 1. Это соотношение между нечетким множеством и его элементом записывается так:

$$\mu(a \in A') \text{ или } \mu_{A'}(a).$$

Теория нечетких множеств (Zadeh, 1973/76; Gaines, Kohout, 1977) еще не построила развернутого аппарата для реализации ФРС. Для осуществления процедуры ФРС приходится использовать традиционный аппарат классической теории множеств и алгебры отношений. Поэтому при построении алгоритма ФРС нечеткие десигнаты ЛЕ и их размытые совокупности необходимо представить в виде четких математических объектов, которые группируются в «классические» множества, имеющие ясно обозначенные границы. Поэтому при формализации десигнативных значений слов, словоформ и словосочетаний приходится однозначно задавать их точные границы.

Эта задача решается с помощью операции контрастной интенсификации (Zadeh, 1973/1974, с. 18) или по другой терминологии — операции сгущения (см.: Пиотровский, 1979, с. 50). Ее существо состоит в следующем:

выбирается пороговое значение m для коэффициента принадлежности μ;

все значения $\mu_{A'}(a) \geqslant m$ увеличиваются, а все значения $\mu_{A'}(a) < m$ уменьшаются, в результате чего нечеткость множества A' сокращается.

При полном сгущении A' все значения $\mu_{A'}(a) \geqslant m$ увеличиваются до единицы, а все $\mu_{A'}(a) < m$ считаются равными нулю, в результате чего нечеткое множество A' превращается в четкое множество A:

$$A = \begin{cases} a \in A, & \text{если } \mu_{A'}(a) \geqslant m \\ a \overline{\in} A, & \text{если } \mu_{A'}(a) < m \end{cases}.$$

В настоящее время не существует достаточно объективной и строгой процедуры задания оценок μ для элементов нечеткого лингвистического множества. Величина μ может быть оценена

статистически или на основании субъективной вероятности, она может быть выведена с помощью экспертных оценок или исходя из структурных соображений. Поэтому операция полного сгущения в инженерно-лингвистической практике осуществляется чаще всего на основании опыта и профессиональной интуиции лингвиста-алгоритмизатора.

При всем этом следует помнить, что, хотя всякое сгущение нечетких множеств связано с потерей информации,[1] обойтись без него при решении нашей задачи не представляется возможным, поскольку, как уже говорилось, при современном состоянии науки алгоритмы ФРС строятся на основе использования четких множеств. Задача алгоритмизатора состоит в том, чтобы, пользуясь традиционно-лингвистическими, лингвостатистическими и другими приемами, максимально сократить при этом потери информации.

3.2. Четко-множественная модель МП

Предполагая, что операция сгущения нечетких лингвистических множеств завершена, перейдем теперь к рассмотрению четко-множественной модели решаемой нами задачи. Прежде всего напомним необходимые теоретико-множественные понятия. Пусть A_1, A_2, . . ., A_n (где n — натуральное число) — некоторые четкие множества. Декартовым произведением множеств A_1, A_2, . . ., A_n называется множество $A_1 \times A_2 \times \ldots \times A_n$, состоящее из всевозможных n-ок $(a_1, a_2, \ldots, a_n)$ $(a_i \in A_i)$, причем $(a_1^{(1)}, \ldots, a_n^{(1)}) = (a_1^{(2)}, \ldots, a_n^{(2)})$ тогда и только тогда, когда для всякого $i \in \{1, 2, \ldots n\}$ имеет место $a_i^{(1)} = a_i^{(2)}$. В случае, когда $n=2$, говорят о декартовом произведении двух множеств и обозначают полученное множество $A_1 \times A_2$.

Пусть $F \subset A \times B$, где A и B — некоторые множества, тогда F называется бинарным отношением между элементами множеств A и B. В случае, когда $A=B$, говорят о бинарном отношении F между элементами множества A или просто о бинарном отношении в A.

Бинарное отношение F между множествами A и B называется однозначным слева, если из того, что пары (a_1, b_1), $(a_2, b_1) \in F$, следует, что $a_1 = a_2$ (инъекция).

Бинарное отношение F между множествами A и B называется однозначным справа, если из того, что (a_1, b_1), $(a_1, b_2) \in F$, следует, что $b_1 = b_2$.

[1] Мера нечеткости множества, аналогичная энтропии, определена в работах: De Luca, Termini, 1972, с. 301—312; Пиотровский, Бектаев, Пиотровская, 1977, с. 361. Эта мера может быть использована для оценки тех информационных потерь, которые возникают при переходе от нечетких множеств к множествам четким.

Бинарное отношение F между множествами A и B называется определенным справа, если для всякого $b \in B$ существует $a \in A$, такое, что $(a, b) \in F$ (сюръекция).

Бинарное отношение F между множествами A и B называется определенным слева, если для всякого $a \in A$ существует $b \in B$, такое, что $(a, b) \in F$.

Бинарное отношение F между элементами множеств A и B называется отображением множества A в множество B, если F однозначно справа и определенно слева (биекция). Отображение F называется взаимно однозначным отображением A на B, если F еще и однозначно слева и определенно справа.

Опираясь на эти условия, обратимся к построению теоретико-множественной модели МП. Начнем с пословно-подстрочного *перевода*.

Пусть A — множество слов данной узкой подобласти входного языка, B — множество соответствующих простых терминов той же узкой подобласти выходного языка, на который производится перевод. Обозначим через $\mathfrak{A}^{(0)} = \{A, B, \theta_i^j\}$ систему двуязычных отношений между лексикой A входного языка и словарем B выходного языка. [Здесь θ_i^j — бинарные отношения между множествами

$$\underbrace{A \times A \times \ldots \times A}_{i\text{-раз}} \text{ и } \underbrace{B \times B \times \ldots \times B}_{j\text{-раз}},$$

возникающие в результате перевода i-сложных входных терминов γ-сложными выходными терминами. Система $\mathfrak{A}^{(0)}$ воплощает отношения входного и выходного словников во всей сложности присущих им лексических, морфологических и синтаксико-семантических деталей. Эти отношения характеризуются колоссальной двуязычной неоднозначностью ЛЕ из A, $\underbrace{A \times \ldots \times A}_{j\text{-раз}}$ относительно B. Алгоритмизировать эти отношения и запрограммировать их на ЭВМ пока не удается. Поэтому необходимо создать некоторую упрощенную модель системы, которая могла бы быть реализована в современном лингвистическом автомате. Рассмотрим один из возможных вариантов такой упрощенной модели.

Используя статистико-дистрибутивные приемы, интуицию носителей двуязычной ситуации (Пиотровский, Чижаковский, 1971, с. 444—451), данные словарей (совокупность этих приемов мы назовем фактором St), проведем упрощение системы $\mathfrak{A}^{(0)}$. Это упрощение достигается путем отбрасывания таких эквивалентов для входных единиц A, $\underbrace{A \times \ldots \times A}_{j\text{-раз}}$, которые имеют низкий коэффициент принадлежности к данному текстовому множеству (т. е. редко встречаются в данном подъязыке) и несут малую смысловую нагрузку.

В результате отбрасывания части отношений θ_i^j система $\mathfrak{A}^{(0)}$ переходит в систему $\mathfrak{A}^{(1)}=\{A, B, \varphi_i^j\}$, где φ_i^j — некоторые из *отношений системы* $\mathfrak{A}^{(0)}$. *Бинарные отношения* φ_i^j в $\mathfrak{A}^{(1)}$ распадаются на две группы:

в первую входят отношения $\varphi_i'^j$, являющиеся отображением A', $A'\times\ \ldots\ \times A'$ в B', $B'\times\ \ldots\times B'$ (это означает, что каждая входная ЛЕ имеет здесь только один переводной эквивалент);

во вторую попадают бинарные отношения $\varphi_i''^j$, которые не являются отображениями в силу их неоднозначности справа, поскольку многие термины из A', $A'\times\ \ldots\ \times A'$ могут переводиться неоднозначно.

Для перевода ЛЕ, охватываемых отношениями $\varphi_i'^j$, используется автоматический словарь слов и оборотов, дающий пословно-пооборотный перевод (Вертель и др., 1971; Автоматическая переработка . . ., 1978). Однако этот АС не устраняет неоднозначности тех входных слов и словоформ, которые характеризуются отношениями $\varphi_i''^j$. Чтобы снять эту неоднозначность, приходится строить специальные алгоритмы, опирающиеся на анализ контекстного окружения многозначных ЛЕ (будем обозначать эти алгоритмы символом Al). Если ввести Al в систему $\mathfrak{A}^{(1)}$, то последняя переходит в систему $\mathfrak{A}^{(2)}=\{A, B, \varphi_i'^j, \psi_i^j\}$. Здесь ψ_i^j суть отображения множеств $A\times\ \ldots\ \times A$ во множества $B\times\ \ldots\ \times B$, полученные из $\varphi_i''^j$ путем снятия неоднозначности с помощью Al.

3.3. Модель устранения многозначности

Рассмотрим математическую постановку задачи снятия многозначности в системе $\mathfrak{A}^{(2)}$, учитывая при этом не только парадигматический, но и синтагматический ее аспекты.

Будем рассматривать входной подъязык, на переработку которого ориентирована наша система МП, как совокупность текстов, образующую единый макротекст фиксированной длины. Тогда, подходя к нему с лексико-семантической точки зрения, можно рассматривать этот текст как множество T, элементами которого являются словоупотребления a_j (эти словоупотребления одновременно могут являться элементами множеств A_1, A_2, . . ., A_n, которые мы рассматривали выше). Сформируем множество U выходных эквивалентов словоформ (словоупотреблений) a_j. *Элементами* U *являются выходные словоформы* b_i.

В ходе лексического МП каждому входному словоупотреблению $a_j \in T$ сопоставляется его выходной эквивалент $b_i \in U$. Это значит, что в основе этого МП лежит соответствие $\varphi : T \to U$, где T — область отправления соответствия, а U — область его прибытия. Элементы области отправления $a_j \in T$ будут по соответствию φ прообразами элементов $b_i \in U$, а элементы $b_i \in U$ из области прибытия являются образами элементов $a_j \in T$ (Шрейдер, 1971, с. 21).

Соответствие $\varphi : T \to U$ *обладает следующими свойствами:*

1) оно всюду определенно, поскольку каждому $a_j \in T$ соответствует хотя бы один его образ $b_i \in U$;[2]

2) это соответствие сюръективно, так как каждому $b_i \in U$ соответствует хотя бы один прообраз a_j из T;

3) соответствие не инъективно, поскольку одному и тому же элементу $b_i \in U$ может соответствовать не один, а *несколько его прообразов из* T;

4) соответствие $\varphi : T \to U$ не является отображением, поскольку одному и тому же $a_j \in T$ может соответствовать не только один, но и несколько образов $b_i \in U$ (ср. рис. 1).

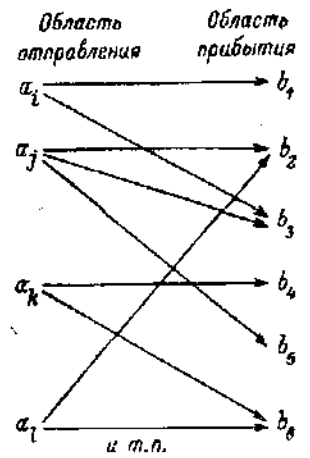


Рис. 1. Соотношение входных словоупотреблений и их эквивалентов при лексическом МП (соотношение φ: $T \to U$).

Задача снятия многозначности сводится к поиску условий, при которых соответствие $\varphi : T \to U$, оставаясь всюду определенным (или не всюду определенным — см. сноску 2), сюръективным и не инъективным, превращалось бы в отображение $\psi : T \to U$ (см. выше).

Для решения этой задачи необходимо перейти с лексического уровня (т. е. от прямого соотнесения входных словоупотреблений и выходных словоформ) на семантический уровень рассуждений (т. е. к формальному анализу значений). При этом словарные *значения многозначных входных словоформ предстают в виде* четких множеств $\hat{a}_j$:

$$\hat{a}_j = \{\hat{a}_j^1, \hat{a}_j^2, \ldots, \hat{a}_j^k\},$$

где $\hat{a}_j^1$, $\hat{a}_j^2$, . . ., $\hat{a}_j^k$, являясь контекстными значениями, выступают в качестве элементов $\hat{a}_j$. Каждому из этих элементов должен быть сопоставлен только один выходной элемент b_i. Формальным воплощением контекстных значений $\hat{a}_j^t$ во входном тексте является комбинаторика входного a_j и его текстового окружения.

Если только что описанный переход проведен, то соотношение $\varphi : T \to U$ превращается в отображение $\psi : T \to U$ показанное на рис. 2а.

[2] Здесь речь идет о МП, *который производится* на базе АС, включающего переводные эквиваленты для всех словоупотреблений входных текстов. В общем случае соответствие φ может не быть всюду определенным. Впрочем, АС, допуская пополнение, позволяет добиться всюду определенности соответствия φ, что и дает нам право считать это соответствие всюду определенным.

Существует два алгоритмических приема перехода от соотношения $\varphi : T \to U$ к отображению $\phi : T \to U$.

Существо первого спискового подхода (см.: Марчук, 1973; Трибис, 1973) состоит в том, что при каждой входной словоформе в АС задаются списки контекстных окружений, которые дают возможность лингвистическому автомату отождествить словоупотребление (с/у) a_j с одним и только с одним из его частных зна-

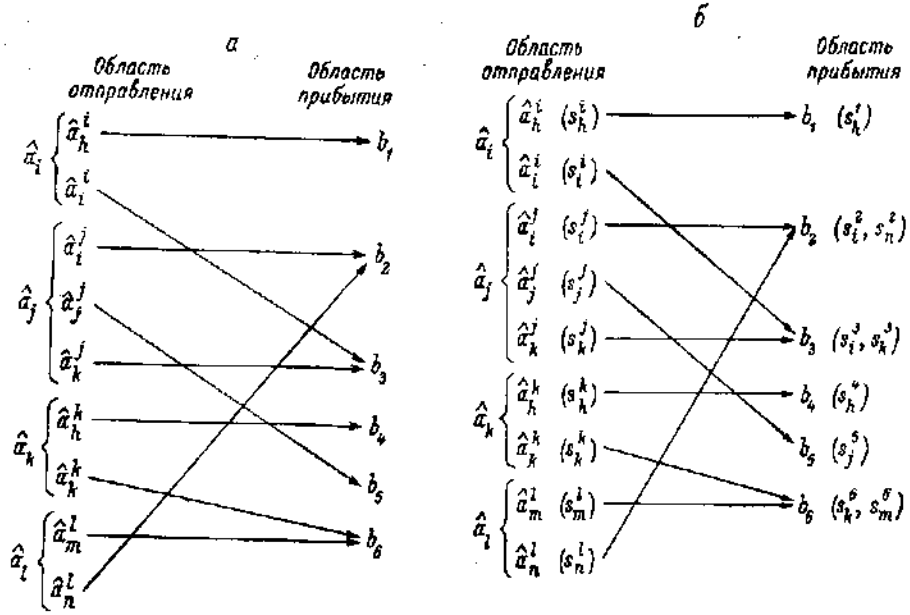


Рис. 2. Отображение ϕ: $T \to U$ при лексико-семантическом МП.

чений $\hat{a}^i_j$. К сожалению, этот подход делает АС слишком громоздким и поэтому оказывается нетехнологичным.

Минимизировать процедуру снятия многозначности удается с помощью второго лексико-семантического подхода. Здесь словарная многозначность входного словоупотребления (с/у) описывается в АС с помощью небольшого набора кодов, каждый из которых соотносит это значение с определенным семантическим классом. Значения выходных эквивалентов также соотнесены с помощью кодов с семантическими классами. Идея устранения многозначности состоит в том, что все сочетания семантических кодов, представляющие в реальном тексте входную цепочку $\ldots a_i, a_j, a_k \ldots$, сопоставляются с заранее заданной таблицей разрешенных сочетаний кодов, которая выступает в роли фильтра, отбрасывающего ложные для данного контекста цепочки кодов. Прошедшая через фильтр цепочка указывает для каждого с/у из последовательности $\ldots a_i a_j a_k \ldots$ тот

семантический код, с помощью которого осуществляется выбор для каждого с/у a_j единственного соответствующего ему в данном контексте переводного эквивалента b_t.

Вся эта лексико-семантическая процедура, реализуемая с помощью семантического автоматического словаря и компактного алгоритма, является, с одной стороны, гораздо более экономной, чем списковый подход, с точки зрения использования объема памяти ЭВМ, а с другой — более гибкой и технологичной при включении новых лексических единиц в АС и систему ФРС.

3.4. Тезаурус

Прежде чем приступить к построению семантического АС и алгоритмического фильтра (их описание см. в 3.5), необходимо выделить семантические признаки, представляющие собой четкие лингвистические объекты и организовать их в виде некоторой четкой системы. Такая организация позволяет выработать для всех $\hat{a}^k_j$ коды, необходимые для построения семантического АС и алгоритма распознавания. Выделение СемП'ов представляет собой лингвистическую процедуру, построенную на операции сгущения нечетких множеств и объектов. Эта процедура будет рассмотрена в § 4. Задачей настоящего раздела является описание построения системы СемП'ов, которая формируется исходя из следующих соображений.

Будем рассматривать входной макротекст T не как множество с/у (см. выше), но как совокупность семантических признаков, присущих с/ф, составляющим этот текст. Зададим соответствие $\beta : T' \to \Theta$, где T' — совокупность СемП'ов реальных словоупотреблений, а Θ — тезаурусная семантическая сеть, включающая классы, т. е. смысловые области (подобласти, микрообласти и т. д.) $S_1, S_2, \ldots, S_k$, соответствующие СемП'ам, и отношения ρ, которые существуют между этими классами. Каждому классу S_i припишем свой код s_i.

Одновременно относительно каждой отдельно взятой входной словоформы имеем

$$\Theta = \bigcup_i S_i \quad \text{и} \quad S_i \cap S_j \neq \varnothing.$$

Соответствие $\beta : T' \to \Theta$ обладает следующими свойствами:

1) оно всюду определенно, поскольку каждому характеризующему входное с/у a_j СемП'у $s_i\,(a_j) \in T'$ соответствует хотя бы один его образ $S_i \in \Theta$;

2) это соответствие сюръективно, так как каждому $S_i \in \Theta$ соответствует хотя бы один прообраз $s_i\,(a_j)$ из T';

3) соответствие β не инъективно, поскольку одному и тому же элементу $S_i \in \Theta$ может соответствовать не один, а несколько его прообразов из T';

4) соответствие $\beta : T' \to \Theta$ является отображением, поскольку одному и тому же $s_i(a_j)$ может соответствовать только один образ $S_i \in \Theta$.[3]

Теперь займемся организацией нашей семантической сети, помня о том, что она должна:

служить средством семантического кодирования каждой входной с/ф;

минимизировать процедуру выработки для каждого класса такого кода, который достаточно прозрачно отражал бы его положение в сети.

Сети, с помощью которых моделируются сложные лингвистические совокупности, представляют собой множества N, элементы которых связаны между одним или более бинарными отношениями. В зависимости от количества отношений, заданных в сети N, ей приписываются ранги.

Так, сетью нулевого ранга

$$N_0 = A = \{x_1, \ldots, x_n\}$$

считается множество лингвистических объектов без каких-либо отношений.

Сетью первого ранга

$$N_1 = \{A, \rho\}$$

будет множество с одним бинарным отношением ρ.

Сеть ранга r

$$N_r = \{A, \rho_1, \rho_2, \ldots, \rho_r\}$$

есть множество с заданными на нем r-отношениями.

Среди сетей ранга r выделяются древесные графы (тезаурусы), характеризующиеся следующими особенностями.

1. Дерево представляется в виде упорядоченного множества, состоящего из одного и более узлов, таких, что имеется один специально обозначенный узел, называемый корнем данного дерева.

2. Остальные узлы (кроме корня) содержатся в $l \geqslant 0$ попарно не пересекающихся множествах $A_1, A_2, \ldots, A_l$, каждое из которых является деревом, деревья $A_1, A_2, \ldots, A_l$ называются поддеревьями данного корня.

3. Узел, не имеющий потомков, называется терминальным узлом (концевой точкой).

4. Заданные на тезаурусе бинарные отношения ρ обладают признаками:

транзитивности, согласно которому, если $(x_i, x_j) \in \rho$ и $(x_j, x_k) \in \rho$, то $(x_i, x_k) \in \rho$, и которое позволяет, таким образом, двигаться по дереву в определенном направлении;

антисимметричности, по которому, если $(x_i, x_j) \in \rho$ и $x_j \neq x_i$, то $(x_j, x_i) \notin \rho$, и которое, таким образом, запрещает двигаться по тезаурусу в направлении, отличном от заданного;

антирефлексивности, исходя из которого не существует $x_i \in A$, такого, что $(x_i, x_j) \in \rho$, и которое не позволяет, двигаясь по тезаурусу, дважды проходить одну и ту же вершину (образовывать петли).

В итоге математическую модель тезауруса можно охарактеризовать следующим образом: пусть A — лингвистическое множество, ρ_A^* — множество бинарных отношений между элементами множества A, определенное лингвистической структурой, тогда тезаурусом называется тройка (A, r, f), где A — множество, r — натуральное число, указывающее на ранг тезаурусной сети (тезауруса), f — отображение множества $M_r = \{1, 2, \ldots, r\}$ во множество ρ_A^*, причем для всякого $i \in M_r$ отношение $f(i) = \rho_{(i)}$ транзитивно, антисимметрично и антирефлексивно.

При построении семантической сети, моделирующей различные предметные области (ПО «именных» понятий, ПО «адъективных» понятий, ПО «глагольных» понятий, ПО локально-темпоральных отношений) (см. §§ 4—5), будет использован дихотомический тезаурус первого ранга, характеризующийся бинарным отношением ρ «род — вид» (рис. 5—9). С помощью такого тезауруса удается не только отразить иерархию семантических классов — признаков и минимизировать процедуру выработки их кодов, но также получить такие двоичные кодовые цепочки, построение которых прозрачно отражает положение каждого СемП'а в системе.

3.5. Устранение многозначности через сочетаемость тезаурусных кодов

Как уже говорилось, словарное значение каждой входной (a_j) и выходной (например, b_2) словоформ представляется в виде множества кодов, выработанных с помощью тезауруса. Каждый код, например s_i^j, указывает на соответствующий ему СемП (в нашем случае $s_i(a_j)$), характеризующий определенное контекстное значение (у нас $\hat{a}_i^j$). Одновременно этот код через семантический класс S_i соотносит $\hat{a}_i^j$ с его переводным эквивалентом —

[3] Отметим, что если в качестве элементов класса рассматривать не СемП'ы, но словарные значения входных словоупотреблений a_j, то Θ превратится в систему пересекающихся классов:

$$\Theta = \bigcup_i S_i \text{ и } S_i \cap S_j \neq \emptyset.$$

При этом соответствие $T' \to \Theta$ перестает быть отображением. Эта ситуация в статье не рассматривается.

с/ф b_2, в значении которой заложен СемП s_i (b_2) и соответствующий код s_i^2, также входящие в семантический класс S_i. Иными словами: семантические коды являются инструментом отображения актуальных значений входных с/ф в значения выходных эквивалентов (см. рис. 2б). При таком подходе основная задача в реализации отображения $\varphi : \mathcal{T} \to U$ сводится к отысканию во множестве

$$\hat{a}_j = \{\ldots s_i(a_j),\ s_j(a_j),\ s_k(a_j),\ \ldots\}$$

такого единственно возможного в данном контексте элемента $s_i(a_j)$, от которого через код s_i^j можно автоматически перейти к искомому выходному эквиваленту b_i.

Эта задача решается следующим образом.

Для каждого словосочетания длиной в m словоформ строится декартово произведение $a_1 \times a_2 \times \ldots \times a_j \times \ldots \times a_m$, элементами которого являются цепочки СемП'ов длиной в m элементов каждая.

Так, например, для трехсловного сочетания $a_i a_j a_k$, в котором словарное значение каждого с/у записано в виде множества

$$\hat{a}_i = \{s_h^i,\ s_t^i\}$$
$$\hat{a}_j = \{s_i^j,\ s_j^j,\ s_k^j\}$$
$$\hat{a}_k = \{s_h^k,\ s_k^k\},$$

декартово произведение $\hat{a}_i \times \hat{a}_j \times \hat{a}_k$ может быть представлено в виде матрицы (рис. 3).

	$\hat{a}_i$	$\hat{a}_j$	$\hat{a}_k$
σ_1	s_h^i	s_i^j	s_h^k
σ_2	s_h^i	s_j^j	s_h^k
σ_3	s_h^i	s_k^j	s_h^k
σ_4	s_h^i	s_i^j	s_k^k
σ_5	s_h^i	s_j^j	s_k^k
σ_6	s_h^i	s_k^j	s_k^k
σ_7	s_t^i	s_i^j	s_h^k
σ_8	s_t^i	s_j^j	s_h^k
σ_9	s_t^i	s_k^j	s_h^k
σ_{10}	s_t^i	s_i^j	s_k^k
σ_{11}	s_t^i	s_j^j	s_k^k
σ_{12}	s_t^i	s_k^j	s_k^k

Рис. 3. Матрица тернарных комбинаций кодов, полученная из декартова произведения $\hat{a}_i \times \hat{a}_j \times \hat{a}_k$.

С помощью лингвистических фильтров в декартовом произведении $\hat{a}_i \times \hat{a}_j \times \hat{a}_k$ отыскивается одна разрешенная семантической системой данного языка (точнее языка и ПО) цепочка σ_8, состоящая из кодов s_α^i, s_β^j, s_γ^k, каждый из которых указывает на тот семантический класс, к которому относится актуальное для данного контекста значение соответствующего входного с/у.

Предположим, что такой цепочкой оказалась

$$\sigma_8 = (s_i^i,\ s_j^j,\ s_k^k).$$

Тогда актуальное текстовое значение с/у a_i будет принадлежать семантическому классу S_i, актуальное значение с/у a_j будет относиться к классу S_j, а текстовое значение с/у a_k — к классу S_k.

Пользуясь схемой отображения, показанной на рис. 2б, можно легко убедиться, что в данном контексте между указанными с/у

и их выходными эквивалентами существует следующее однозначное соотношение:

$$a_i \to b_3$$
$$a_j \to b_5$$
$$a_k \to b_4.$$

Подведем итог.

Первым шагом на пути построения алгоритма ФРС текста в двуязычной ситуации является применение статистического огрубления (St) системы входного естественного языка $\mathfrak{A}^{(0)}$ и замены ее упрощенной моделью $\mathfrak{A}^{(1)}$. Однако переход

$$\mathfrak{A}^{(0)} \xrightarrow{\text{St}} \mathfrak{A}^{(1)}$$

не обеспечивает полного распознавания смысла у составляющих входной текст ЛЕ. Хотя модель $\mathfrak{A}^{(1)}$ и является огрублением системы $\mathfrak{A}^{(0)}$, она все же еще сохраняет такую особенность ЕЯ, как многозначность ЛЕ, которая препятствует полной и эффективной семантической переработке естественного языка на ЭВМ.

Следующим шагом на пути построения ФРС является приложение к модели $\mathfrak{A}^{(1)}$ алгоритма устранения многозначности (Al), опирающегося на тезаурусное описание (θ) семантики ЛЕ. В результате перехода

$$\mathfrak{A}^{(1)} \xrightarrow{\text{Al},\ \theta} \mathfrak{A}^{(2)}$$

мы и получаем только что описанную математическую модель устранения многозначности $\mathfrak{A}^{(2)}$.

Прежде чем говорить о реальных возможностях $\mathfrak{A}^{(2)}$ с точки зрения МП (см. § 5), рассмотрим лингвистические приемы отбора СемП'ов и организации тезауруса.

§ 4. Отбор семантических признаков и построение тезауруса

В современной инженерной лингвистике чаще всего применяется три приема выделения СемП'ов:

использование общепринятых логико-смысловых категорий;

анализ словарных дефиниций, дополняемых матричным анализом синонимии;

анализ контекстов.

Все эти приемы используются в настоящей работе.

Исходя из дихотомического принципа построения тезауруса (§ 3), выбранные СемП'ы группируются в пары таким образом, что первый из них указывает на наличие (присутствие) признака (A), а второй — на его отсутствие (не-A). Эти пары мы будем называть дифференциально-семантическими признаками (ДСП).

4.1. Использование логико-смысловых категорий

Известно, что формами существования материи являются пространство и время. Поэтому концептуальная модель (СП) должна включать эти логико-смысловые категории в качестве семантических признаков. В связи с этим в наборе СемП'ов, характеризующих предложно-падежные отношения, предусмотрен СемП «локальность», который в паре с его контрастным признаком «нелокальность» образуют ДСП «локальность — нелокальность» (in the house — *during the day, à la maison — *pendant le jour, *в доме — *в течение дня*).

Аналогичным образом возникает ДСП «темпоральность — нетемпоральность» (during the day —*rewarded for bravery, pendant le jour — *décoré pour le courage, *в течение дня — *награжденный за храбрость*).

По этой же схеме выделяются ДСП у существительных, прилагательных и глаголов. Так, например, исходными ДСП существительных становятся пары логико-смысловых признаков «конкретное — неконкретное (абстрактное)» (ср. пары train — *phenomenon, *phénomène, *поезд — *явление*).[4] Для прилагательных такой парой будет ДСП «собственный признак (качественное прилагательное) — несобственный признак (относительное прилагательное)» (ср. green — *political, vert — *politique, *зеленый — *политический*). Для глагола в качестве примера такого ДСП можно указать на пару «отношение — неотношение» (ср. have — *fly, avoir — *voler, *иметь — *летать*).

Логико-смысловые категории и конструируемые на их основе СемП'ы и ДСП «конкретное — неконкретное (абстрактное)» являются обычно нечеткими лингвистическими объектами. Примером тому служат понятия абстрактности и конкретности, которые в философском плане не противопоставляются друг другу, но находятся в тесной связи (Кондаков, 1976). Сгущение и дискретизация этих понятий в виде СемП'ов осуществляется алгоритмизатором интуитивно на основе коллективного опыта носителей языка, обладающих определенными мировоззренческими установками.

4.2. Анализ словарных дефиниций

Источником выделения СемП'ов могут служить также дефиниции толковых словарей и данные словарей синонимов.

[4] Читателя не должен смущать тот факт, что большинство приводимых примеров не являются антонимическими парами. Дело в том, что каждый контрастирующий СемП охватывает все семы, находящиеся в узлах куста, выходящего из данного СемП'а. Поэтому, если для слова, иллюстрирующего СемП *А*, не удается подобрать «чистый» антоним, иллюстрирующий СемП не-*А*, то в этом случае приводится любое слово, у которого отсутствует признак *А*, разумеется, при условии, что оно принадлежит к кусту не-*А*. Такой квазиантоним обозначается повсюду знаком *, стоящим перед квазиантонимом.

Общность словарных дефиниций, основанная на семантической смежности слов одного синонимического ряда, позволяет выделить некоторую инвариантную часть их значения, которая рассматривается в качестве семантического признака.[5]

Рассмотрим несколько примеров.

Толковые словари современного английского языка Уэбстера (Webster, 1971), Хорнби и др. (Hornby, Gatenby, Wakefield, 1958) дают следующие дефиниции элементов синонимического ряда, обладающих общим признаком 'утаивать, скрывать, маскировать':

conceal — keep secret from,
mask — conceal from view,
suppress — keep secret, not reveal,
cover — conceal (feelings, etc.), conceal by wrapping up,
hide — put, keep out of sight, keep secret from, keep from view, conceal,
cloak — conceal, disguise,
shroud — cover, conceal or disguise,
veil — cover with, conceal, disguise, mask.

Это инвариантное значение положено в основу глагольного СемП'а «элиминация объекта», понимаемого как устранение объекта из поля рассмотрения.

Исходя из принципа дихотомического построения тезауруса, одновременно выделяется СемП «неэлиминация объекта». Оба эти признака образуют ДСП «элиминация — неэлиминация» объекта, входящий в СП глагола.

Аналогичным образом выделяются СемП'ы прилагательных. Например, рассмотрение дефиниций, взятых из словаря Ларусса (Grand Larousse, 1971—1972; ср.: Trésor, 1971 и сл.):

local — qui est propre à un lieu déterminé, relatif à une région précise;
régionel — qui comprend une certaine région, propre à une région —

дает возможность выделить СемП «относящийся к месту» и одновременно его антонимический коррелят «неотносящийся к месту», которые образуют ДСП «относящийся — неотносящийся к месту», включающийся в СП прилагательного.

Рассмотрение таких словарных определений в «Словаре современного русского литературного языка» (1948—1965), как:

однотипность — *свойство однотипного, одинаковость, сходство по типу с другим;*

[5] Следует сразу же подчеркнуть, что подбор дефиниций и выбор инварианта всегда определяются прагматикой и осуществляются исходя из конкретной коммуникативной установки лингвиста или конкретного коллектива специалистов, для удовлетворения информационных потребностей которых строится данная система МП и ориентированный на нее тезаурус.

эквивалентность — свойство эквивалентного, равнозначного, равносильного;
четкость — свойство, качество четкого;
эффективность — свойство эффективного,

позволяет выделить СемП «свойство» и ДСП «свойство — несвойство», входящие в СП существительного.

Словарные определения, так же как и логико-смысловые категории, не являются четкими лингвистическими объектами и множествами. Их сгущение и дискретизация в СемП'ы и ДСП производятся лингвистом-алгоритмизатором, исходя из авторитета используемых им лексикографических источников, собственной интуиции и субъективных лингвистических вероятностей.

В тех случаях, когда толковые словари не дают достаточного материала для однозначного выделения четких СемП'ов и ДСП, обращаются к матричному анализу синонимических рядов.

Сущность этого приема состоит в следующем.

Образующие синонимический ряд ЛЕ находятся в отношении взаимного дефинирования. Например, если в качестве синонимов *слова плохой* указываются слова *гадкий*, *нехороший*, *скверный* и другие (СлС, 1975; Александрова, 1975; Клюева, 1956), то последние рассматриваются в качестве семантических компонентов (сем), входящих в значение слова *плохой*. Одновременно если синонимами *скверный* являются *плохой* и *гадкий*, то последние два слова квалифицируются как семы значения слова *скверный*. Эти семантические отношения представляются обычно в виде матрицы (рис. 4), в строках которой указываются слова, входящие в синонимический ряд, а в столбцах — семантические компоненты значений этих слов. Если по данным синонимических словарей два *слова* (a_i *и* a_j) находятся в отношении синонимии, то на пересечении строки, в которой стоит дефинируемое слово (a_i), и столбца, в котором указывается синоним — сема (a_j), ставится плюс (+). Те члены синонимического ряда, которые содержат наибольшее количество сем, рассматриваются в качестве выразителя некоторых общих семантических признаков (ср. *плохой* и *гадкий*). Для этих признаков с помощью синонимических словарей и словарей антонимов (Webster, 1968; Комиссаров, 1964; Bailly, 1967; Dupuis, 1961; СлС, 1975; Александрова, 1975; Клюева, 1956; Львов, 1978) подбираются «анти-СемП'ы», которые, как уже говорилось, вместе с исходными СемП'ами образуют ДСП.

Рассмотрим всю эту процедуру на примере нашего заглавного синонимического ряда.

Каждый синонимический ряд представляет собой обычно нечеткое множество, переходящее благодаря размытости своих границ в другие множества синонимов и образующее вместе с этими множествами бесконечный синонимический макроряд (Пиотровский, 1979, с. 46—47). Поэтому, прежде чем приступить к построению матрицы, необходимо наш размытый синонимический ряд

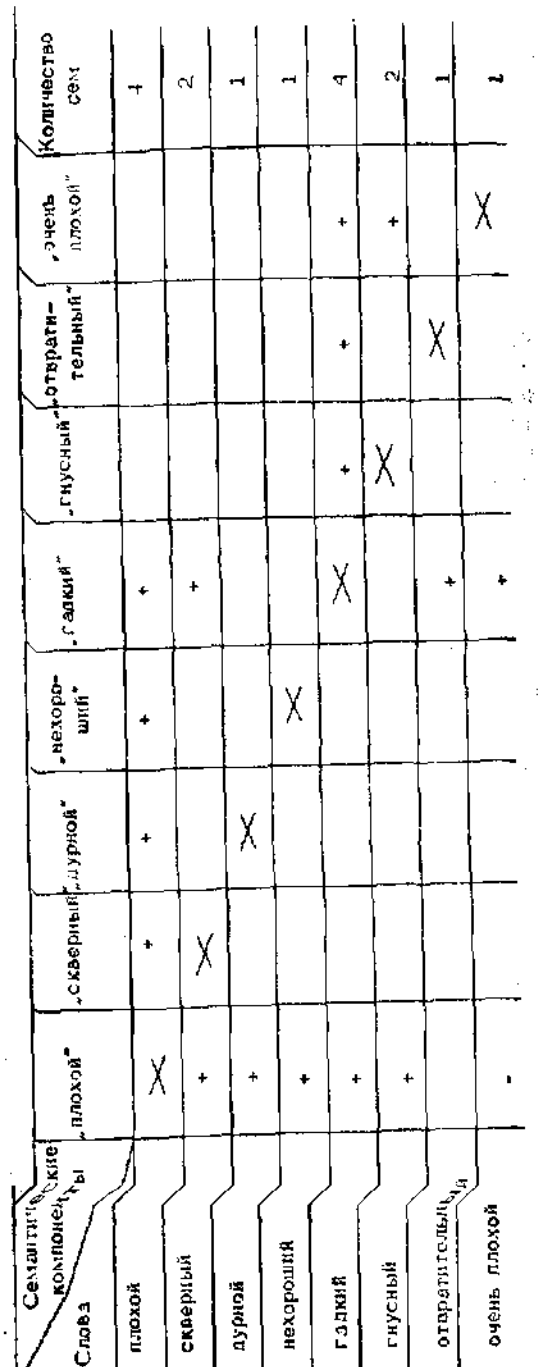


Рис. 4. Матрица компонентной структуры русских прилагательных, принадлежащих к синонимическому ряду с инвариантным значением 'плохой'.

превратить в четкое лингвистическое множество. Эта операция сгущения осуществляется путем сохранения в синонимическом ряду только общеупотребительных слов. Все слова узкостилевого или экспрессивного употребления, имеющие соответствующие пометы, например, *аховый* (прост., усилит.), *дрянной* (прост., пренебр.), *паршивый* (прост., презр.), *поганый* (прост., презр.), *хреновый* (груб.-прост.), *худой* (разг.) и другие, рассматриваются как имеющие малый коэффициент принадлежности к указанному множеству синонимов (§ 3) и поэтому в матрицу не вносятся. После того, как матрица сформирована, можно перейти к ее анализу.

Большим количеством семантических компонентов и поэтому широким значением обладает в нашем синонимическом ряду прилагательное *плохой*. Поэтому оно может рассматриваться в качестве воплощения СемП'а «отрицательная оценка». Соответственно его антоним *хороший* выявляет СемП «положительная оценка». Оба эти СемП'а формируют ДСП «положительная — отрицательная оценка» (табл. 3, рис. 7). Достаточно большое количество сем обнаруживается также в значениях слов *гадкий* и *гнусный*. Это дает возможность рассматривать их в качестве носителей еще одного СемП'а, находящегося на более низком уровне классификации, чем ДСП «положительная — отрицательная оценка», и дающего более тонкую градацию СемП'а «отрицательная оценка». Исходя из семантики прилагательных *гадкий*, *гнусный*, *отвратительный*, *очень плохой*, можно выделить СемП «сильная отрицательная оценка», который вместе с антипризнаком «слабая отрицательная оценка» (ср. *нехороший*) образуют ДСП «сильная — слабая отрицательная оценка». Аналогичным путем с помощью анализа синонимических и антонимического словарей конструируется ДСП «сильная — слабая положительная оценка» (*прекрасный — неплохой*).

Аналогичная процедура применялась и относительно французских синонимических рядов для выделения других ДСП СП прилагательных (Васильева, 1978).

4.3. Выделение семантических признаков путем анализа контекстов

Только что рассмотренные приемы обнаружения семантических признаков и ДСП дают удовлетворительные результаты при выделении именных и отчасти адъективных СемП'ов, однако они не всегда оказываются эффективными при выявлении предложно-падежных и глагольных СемП'ов. Дело в том, что предлог, падежная морфема и отчасти глагольное слово имеют настолько широкое и размытое виртуальное значение, что его не удается сгустить в четкий и дискретный смысловой инвариант. Поэтому для выявления тех смысловых инвариантов, которые могут быть

положены в основу предложно-падежных и глагольных СемП'ов, необходимо поместить предлог (падежную морфему, глагол) в некоторый ограничивающий его значение контекст. В качестве такого минимального контекста выступают обычно правая и отчасти левая дистрибуции указанных слов. Поэтому при выделении указанных СемП'ов широко используется семантический анализ сегментов «глагол (существительное)+предлог+существительное», «глагол+существительное».

В качестве примера рассмотрим сочетания английского многозначного предлога to. В сочетании a letter to Rome этот предлог обнаруживает СемП «конечная граничность», в сочетании to get to the frontier он выявляет СемП «предельность», а в словосочетании a message to the government демонстрирует признак «адресатность».

Аналогичная процедура применяется при выделении некоторых глагольных ДСП.

4.4. Лингвистические принципы построения тезауруса

Все три рассмотренные приема обеспечивают выделение подавляющего количества СемП'ов, необходимых для создания поля признаков, характеризующих наше СП. Однако при системной организации этих признаков они должны быть упорядочены в тезаурусной сети, и здесь возникают две задачи:

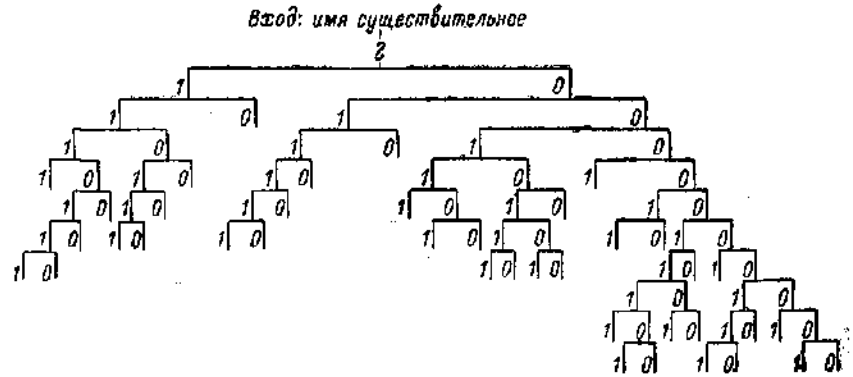


Рис. 5. Тезаурусный граф «Имя существительное».

1) установление иерархии ДСП;

2) выявление отдельных лакун в системе и заполнение их путем дедуктивного задания новых СемП'ов и ДСП.

Иерархия ДСП устанавливается исходя из степени их абстрактности. ДСП, выведенные из общих логико-смысловых категорий, например, «конкретное — неконкретное (абстрактное)», «локальность — нелокальность (темпоральность)», занимают верх-

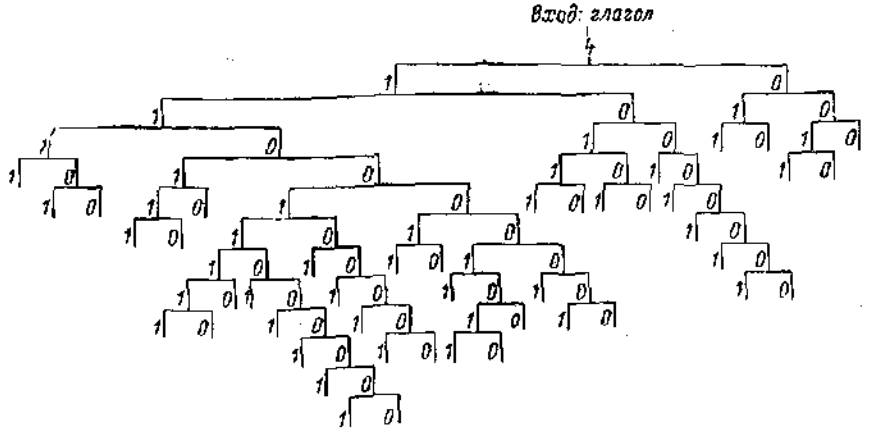


Рис. 6. Тезаурусный граф «Глагол».

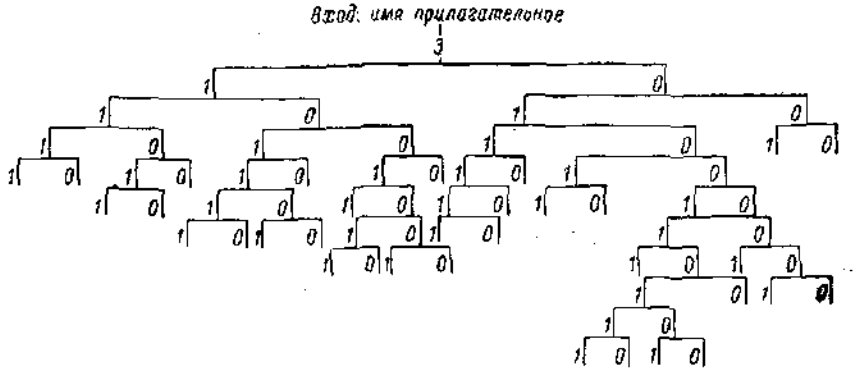
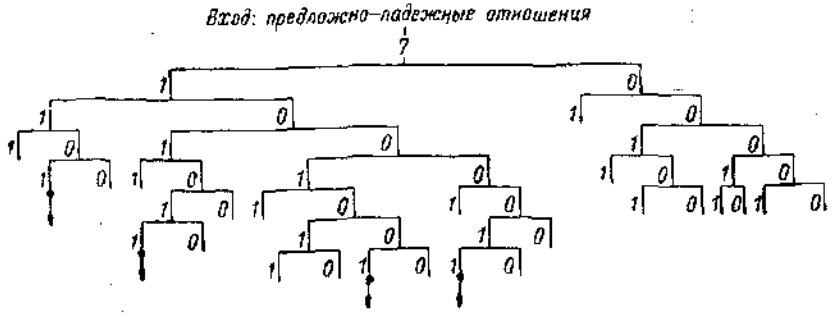


Рис. 7. Тезаурусный граф «Имя прилагательное».



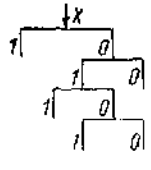


Рис. 8. Тезаурусный граф «Предложно-падежные отношения» (фрагмент).

Примечание. Дополнительный граф, обозначенный буквой X, присоединяется к терминальным узлам основного графа, обозначенным стрелками.

Таблица 1

СемП'ы имен существительных и их коды

Код СемП'а	Название СемП'а	Примеры: английский, французский, русский
2—1	*конкретное*	train, *поезд*
2—0	неконкретное	phenomenon, phénomène, *явление*
2—11	дискретное	train, *поезд*
2—10	недискретное	water, eau, *вода*
2—111	тело	train, *поезд*
2—110	нетело	town, ville, *город*
2—1111	артефакт	train, *поезд*
2—1110	неартефакт	tree, arbre, *дерево*
2—11101	одушевленное	person, personne, *лицо*
2—11100	неодушевленное	object, chose, *предмет*
2—111011	отдельное лицо	president, président, *президент*
2—111010	неотдельное лицо (группа лиц)	youth, jeunesse, *молодежь*
2—1110111	имя собственное	Smith, *Смит*
2—1110110	и. несобственное (нарицательное)	delegate, délégué, *представитель*
	2—110 нетело	
2—1101	место	town, ville, *город*
2—1100	неместо	*
2—11011	административно-географический объект	country, pays, *страна*
2—11010	неадминистративно-географический о.	territory, territoire, *территория*
2—110111	топоним	London, Londres, *Лондон*
2—110110	нетопоним	*
	2—0 неконкретное	
2—01	процесс	development, développement, *развитие*
2—00	непроцесс	colour, couleur, *цвет*
2—011	спецпроцесс	aggression, agression, *агрессия*
2—010	неспецпроцесс	development, développement, *развитие*
2—0111	политико-экономический процесс	trade, commerce, *торговля*
2—0110	неполитико-экономический п.	*
2—01111	торгово - промышленная операция	trade, commerce, *торговля*
2—01110	неторгово-промышленная оп.	strike, grève, *забастовка*
2—011111	финансовая оп.	devaluation, dévaluation, *девальвация*
2—011110	нефинансовая оп.	*
	2—00 непроцесс	
2—001	признак	colour, couleur, *цвет*
2—000	непризнак	state, condition, *состояние*
2—0011	свойство	fluctuation, *колебание*

Таблица 1 (*продолжение*)

Код СемП'а	Название СемП'а	Примеры: английский, французский, русский
2—0010	несвойство	year, an, *год*
2—00111	способность	force, *сила*
2—00110	неспособность	incapacity, inaptitude, *неспособность*
2—001101	возможность	potential, potentiel, *потенциал*
2—001100	невозможность	impossibility, impossibilité, *невозможность*
	2—0010 несвойство	
2—00101	параметр	level, niveau, *уровень*
2—00100	непараметр	*
2—001011	количественный параметр	number, nombre, *число*
2—001010	неколичественный пар.	year, an, *год*
2—0010111	равенство	equivalent, équivalent, *эквивалент*
2—0010110	неравенство	minority, minorité, *меньшинство*
2—0010101	временной параметр	year, an, *год*
2—0010100	невременной пар.	*
	2—000 непризнак	
2—0010	состояние	existence, état, *существование*
2—0000	несостояние	party, partie, *партия*
2—00001	отношение	cooperation, collaboration, *сотрудничество*
2—00000	неотношение	plant, farbique, *завод*
2—000011	общественно-экономическая формация	socialism, socialisme, *социализм*
2—000010	необщественно-экономическая ф.	*
	2—00000 неотношение	
2—000001	институт	monopoly, monopolie, *монополия*
2—000000	неинститут	money, argent, *деньги*
2—0000011	форма организации института	union, *союз*
2—00000010	неформа о. и.	*
2—00000111	внутригосударственный институт	government, gouvernement, *правительство*
2—00000110	невнутригосударственный и.	UNESCO, *ЮНЕСКО*
2—000001101	международный и.	UNESCO, *ЮНЕСКО*
2—000001100	немеждународный и.	*
2—000001111	орган управления	government, gouvernement, *правительство*
2—000001110	неорган у.	party, partie, *партия*
2—0000011101	предприятие	plant, usine, *завод*
2—0000011100	непредприятие (учреждение)	bank, banc, *банк*

Таблица 1 (*продолжение*)

Код СемП'а	Название СемП'а	Примеры: английский, французский, русский
	2—000000 неинститут	
2—0000001	отрасль народного хозяйства	industry, industrie, *промышленность*
2—0000000	неотрасль н. х.	price, prix, *цена*
2—00000001	финансово-экономическая категория	price, prix, *цена*
2—00000000	нефинансово-экономическая к.	newspaper, journal, *газета*
2—000000001	форма информации	report, rapport, *доклад*
2—000000000	неформа и.	opinion, *мнение*
2—0000000001	форма мыслительной деятельности	intention, *намерение*
2—0000000000	неформа м. д.	*
2—000000011	денежный документ	banknote, bank-note, *банкнота*
2—000000010	неденежный д.	bill, projet, *законопроект*
2—0000000111	денежная единица	franc, *франк*
2—0000000110	неденежная е.	metre, mètre, *метр*

Примечание 1. В этой и таблицах 2—4 в третьей графе при совпадении орфографии английского и французского слов дается одно написание. 2. Звездочки в третьей графе означают, что соответствующие СемП'ы не имеют реализации в данном СП.

Таблица 2

СемП'ы глаголов и их коды

Код СемП'а	Название СемП'а	Примеры: английский, французский, русский
4—1	отношение	have, avoir, *иметь*
4—0	неотношение	be, être, *быть*
4—11	операциональность	make, faire, *делать*
4—10	неоперациональность	have, avoir, *иметь*
4—111	перемещение объекта	carry, transporter, *перевозить*
4—110	неперемещение о.	be, se trouver, *находиться*
4—1111	передача объекта лицу	send, passer, *посылать*
4—1110	непередача о.	receive, recevoir, *получать*
4—11101	отчуждение о.	take, prendre, *взять*
4—11100	неотчуждение о. (уступительность)	give, donner, *давать*
	4—110 неперемещение	
4—1101	производство материального объекта	build, construire, *строить*
4—1100	непроизводство м. о.	process, cultiver, *обрабатывать*
4—11011	конструктивность идеального объекта	organize, organiser, *организовывать*
4—11010	неконструктивность и. о.	*
4—110111	генерация объекта	generate, causer, *порождать*
4—110110	негенерация о.	*

Таблица 2 (*продолжение*)

Код СемП'а	Название СемП'а	Примеры: английский, французский, русский
	4—1100 непроизводство	
4—11001	*обработка объекта*	dry, sécher, *сушить*
4—11000	необработка о.	control, contrôler, *контролировать*
4—110011	модификация о.	change, changer, *изменять*
4—110010	немодификация о.	compare, comparer, *сравнивать*
4—1100111	физическое воздействие	cut, coupre, *резать*
4—1100110	нефизическое в.	aggravate, aggraver, *обострять*
4—11001111	*поражающее в.*	beat, battre, *бить*
4—11001110	непоражающее в.	*
4—110011111	уничтожение объекта	destroy, détruire, *разрушать*
4—110011110	неуничтожение о.	*
	4—1100110 нефизическое воздействие	
4—11001101	увеличение объекта	extend, élargir, *расширять*
4—11001100	неувеличение о.	limit, limiter, *ограничивать*
4—110011001	уменьшение о.	reduce, diminuer, *сокращать*
4—110011000	неуменьшение о.	develop, développer, *развивать*
4—1100110001	консолидация о.	consolidate, consolider, *укреплять*
4—1100110000	неконсолидация о.	aggravate, empirer, *ухудшать*
4—11001100001	элиминация о.	cancel, annuler, *отменять*
4—11001100000	неэлиминация о.	speed up, presser, *торопить*
4—1100110000001	ускорение о.	accelerate, précipiter, *ускорять*
4—1100110000000	неускорение (торможение) о.	slowdown, freiner, *тормозить*
	4—110010 немодификация	
4—1100101	идентификация объекта	identify, identifier, *отождествлять*
4—1100100	неидентификация о.	search, chercher, *искать*
4—11001001	избирательность о.	select, choisir, *выбирать*
4—11001000	неизбирательность о.	join, joindre, *присоединять*
4—110010001	объединение о.	combine, unir, *объединять*
4—110010000	необъединение (разъединение) о.	separate, désunir, *разъединять*
4—1100100001	классификация о.	sort, trier, *сортировать*
4—1100100000	неклассификация о.	expose, exposer *выставлять*
4—11001000001	экспозитивность о.	display, montrer, *показывать*
4—11001000000	неэкспозитивность (изоляция) о.	conceal, cacher, *прятать*

Таблица 2 (*продолжение*)

Код СемП'а	Название СемП'а	Примеры: английский, французский, русский
	4—1100 необработка	
4—110001	противодействие объекту	oppose, empêcher, *препятствовать*
4—110000	непротиводействие (содействие) о.	promote, aider, *содействовать*
4—1100011	агрессивное воздействие	attack, attaquer, *нападать*
4—1100010	неагрессивное в.	defend, défendre, *защищать*
	4—110000 непротиводействие	
4—1100001	управление объектом	govern, diriger, *руководить*
4—1100000	неуправление о.	fulfil, accomplir, *выполнять*
4—11000011	директивность	direct, diriger, *направлять*
5—11000010	недирективность	decide, résoudre, *решать*
4—110000101	санкция	dismiss, débaucher, *увольнять*
4—110000100	несанкция	*
4—1100001011	резолютивная акция	confirm, sanctionner, *утвердить*
4—1100001010	нерезолютивная а.	*
	4—1100000 неуправление	
4—11000001	торгово-финансовая акция	purchase, acheter, *покупать*
4—11000000	неторгово-финансовая а.	apply, appliquer, *применять*
4—110000001	утилизация объекта	use, profiter, *использовать*
4—110000000	неутилизация о.	*
	4—10 неоперациональность	
4—101	экзистенциональность	visit, fréquenter, *посещать*
4—100	неэкзистенциональность	want, vouloir, *хотеть*
4—1011	темпоральность	precede, précéder, *предшествовать*
4—1010	нетемпоральность	contain, contenir, *содержать*
4—10111	инициальность	begin, commencer, *начинать*
4—10110	неинициальность (терминальность)	finish, finir, *заканчивать*
4—10101	квалификативность	possess, posséder, *обладать*
4—10100	неквалификативность (квантитативность)	enumerate, compter, *исчислять*
	4—100 неэкзистенциональность	
4—1001	мыслительная деятельность	understand, comprendre, *понимать*
4—1000	немыслительная д.	hate, haïr, *ненавидеть*
4—10001	эмоциональная д.	feel, sentir, *чувствовать*
4—10000	неэмоциональная д.	examine, examiner, *рассматривать*
4—100001	зрительная д.	inspect, inspecter, *осматривать*
4—100000	незрительная д.	tell, parler, *рассказывать*
4—10000001	информативная д.	inform, informer, *сообщать*
4—10000000	неинформативная д.	*

Таблица 2 (*продолжение*)

Код СемП'а	Название СемП'а	Примеры: английский, французский, русский
	4—0 неотношение	
4—01	состояние субъекта	shiver, geler, *мерзнуть*
4—00	несостояние с.	run, courir, *бежать*
4—011	процесс	die, mourir, *умирать*
4—010	непроцесс	*
4—001	свойство субъекта	fly, voler, *летать*
4—000	несвойство с.	*
4—0011	функциональность с.	work, travailler, *работать*
4—0010	нефункциональность с.	*

Таблица 3

СемП'ы имен прилагательных и их коды

Код СемП'а	Название СемП'а	Примеры: английский, французский, русский
3—1	собственно признак	green, vert, *зеленый*
3—0	несобственно пр.	political, polytique, *политический*
3—11	оценочный пр.	nice, beau, *хороший*
3—10	неоценочный пр.	wide, large, *широкий*
3—111	положительная оценка	nice, beau, *хороший*
3—110	неположительная оц.	bad, mauvais, *плохой*
3—1111	сильная положительная оц.	excellent, *прекрасный*
3—1110	несильная положительная оц.	pretty good, assez, *неплохой*
3—1101	отрицательная оц.	bad, mauvais, *плохой*
3—1100	неотрицательная оц.	*
3—11011	сильная отрицательная оц.	disgusting, détestable, *гадкий*
3—11010	несильная отрицательная оц.	bad, mauvais, *нехороший*
	3—10 неоценочный признак	
3—101	параметрический признак	wide, large, *широкий*
3—100	непараметрический пр.	angry, méchant, *злой*
3—1011	размерный	wide, large, *широкий*
3—1010	неразмерный	*
3—10111	целый	total, entier, *целый*
3—10110	нецелый	partial, partiel, *частичный*
3—101111	большой (целый)	large, grand, *большой*
3—101110	небольшой (целый)	little, petit, *малый*
3—101101	большой (нецелый)	large, grand, *большой*
3—101100	небольшой (нецелый)	little, petit, *маленький*
	3—100 непараметрический	
3—1001	обозначающий состояние	ill, malade, *больной*
3—1000	необозначающий с.	*
3—10011	физическое с.	solid, dur, *твердый*

Таблица 3 (*продолжение*)

Код СемП'а	Название СемП'а	Примеры: английский, французский, русский
3—10010	нефизическое с.	nervous, nerveux, *нервный*
3—100101	психическое с.	nervous, nerveux, *нервный*
3—100100	непсихическое с.	rich, riche, *богатый*
3—1001001	экономическое с.	rich, riche, *богатый*
3—1001000	неэкономическое с.	*
3—1001011	эмоциональное с.	angry, méchant, *злой*
3—1001010	неэмоциональное с.	*
	3—0 несобственно признак	
3—01	обозначающий отношения	political, polytique, *политический*
3—00	необозначающий отн.	pacific, pacifique, *мирный*
3—011	принадлежащий кому-либо, чему-либо	personal, personnel, *личный*
3—010	непринадлежащий кому-либо, чему-либо	habitual, habituel, *обычный*
3—0111	принадлежащий лицу	personal, personnel, *личный*
3—0110	непринадлежащий л.	*
3—01111	принадлежащий группе лиц	popular, populaire, *народный*
3—01110	непринадлежащий гр. л.	*
3—011111	национальный	English, anglais, *английский*
3—011110	ненациональный	*
	3—010 непринадлежащий кому-либо или чему-либо	
3—0101	относящийся ко времени	habitual, habituel, *обычный*
3—0100	неотносящийся ко вр.	territorial, *территориальный*
3—01011	прошедший	ancient, ancien, *древний*
3—01010	непрошедший	modern, moderne, *современный*
	3—0100 неотносящийся ко времени	
3—01001	относящийся к месту	local, *локальный*
3—01000	неотносящийся к м.	numerous, nombreux, *многочисленный*
3—010001	относящийся к количеству	numerous, nombreux, *многочисленный*
3—010000	неотносящийся к кол.	social, *социальный*
3—0100011	единичный	unique, *единственный*
3—0100010	неединичный	double, *двойной*
	3—010000 неотносящийся к количеству	
3—0100001	относящийся к социальному устройству общества	democratic, démocratique, *демократический*
3—0100000	неотносящийся к соц. устр. общ.	monetary, monétaire, *денежный*
3—01000011	относящийся к организации общ.	parliamentary, parlementaire, *парламентский*
3—01000010	неотносящийся к орг. общ.	*
3—010000111	относящийся к отрасли народного хозяйства	industrial, industriel, *промышленный*

Таблица 3 (*продолжение*)

Код СемП'а	Название СемП'а	Примеры: английский, французский, русский
3—010000110	неотносящийся к нар. хоз.	electoral, électoral, *выборный*
3—0100001111	относящийся к науке	scientific, scientifique, *научный*
3—0100001110	неотносящийся к н.	*
3—0100001101	относящийся к социальным мероприятиям	electoral, électoral, *выборный*
3—0100001100	неотносящийся к соц. мер.	*
3—01000001	относящийся к финансово-экономической категории	monetary, monetaire, *денежный*
3—01000000	неотносящийся к фин.-экон. кат.	

3—00 необозначающий отношения

3—001	обозначающий состояние	pacific, pacifique, *мирный*
3—000	необозначающий сост.	*

Таблица 4

Предложно-падежные СемП'ы и их коды

Код СемП'а	Название СемП'а	Примеры: английский, французский, русский
7—1	локальность	in the house, à la maison, *в доме*
7—0	нелокальность	during the day, pendant le jour, *в течение дня*

7—1 локальность

7—11	отношение к точке (к нульмерному пространству)	at the point, au point, *в точке*
7—10	о. не к точке (не к нульмерному пр.)	in the house, à la maison, *в доме*
7—101	о. к линии (к одномерному пр.)	on the line, sur la ligne, *на линии*
7—100	о. не к линии (не к одномерному пр.)	in the house, à la maison, *в доме*
7—1001	о. к поверхности (к двухмерному пр.)	on the floor, sur le plancher, *на полу*
7—1000	о. не к поверхности (т. е. к трехмерному пр.)	in the house, à la maison, *в доме*
7—111 7—1011 7—10011 7—10001	преодоление пространства	through the field, à travers le champ, *через поле*
7—110 7—1010 7—10010 7—10000	непреодоление пр.	in the field, au champ, *в поле*

Таблица 4 (*продолжение*)

Код СемП'а	Название СемП'а	Примеры: английский, французский, русский
7—10101 7—100101 7—100001	отношение к замкнутому пр. *	in the arena, dans les limites de l'arène, *в пределах арены*
7—10100 7—100100 7—100000	о. не к замкнутому пр.	on the arena, sur l'arène, *на арене*
7—1101 7—101011 7—1001011 7—1001001 7—1000011	нахождение или направленность в пределах (в пределы) пр.	in the room, dans la pièce, *в комнате*
7—1100 7—101010 7—1001010 7—1001000 7—1000010	нахождение или направленность вне (за пределы) пр.	outside the room, en dehors de la pièce, *вне (за пределами) комнаты*
7—X 1	статика	stay in the house, rester à la maison, *оставаться в доме*
7—X 0	нестатика (динамика)	penetrate into the house, pénétrer dans la maison, *проникнуть в дом*
7—X 01	граничность	a letter from Paris, la lettre de Paris, *письмо из Парижа*
7—X 00	неграничность	travel on the channel, aller le long du canal, *ехать по каналу*
7—X 011	начальная граничность	a letter from Paris, la lettre de Paris, *письмо из Парижа*
7—X 010	неначальная (конечная) гр.	a letter to Rome, la lettre à Rome, *письмо в Рим*
7—X 0101	предельность	get to the frontier, aller jusqu'à la frontière, *добраться до границы*
7—X 0100	непредельность	move towards the frontier, se diriger vers la frontière, *двигаться к границе*

7—0 нелокальность

7—01	темпоральность	during the day, pendant le jour, *в течение дня*
7—00	нетемпоральность	rewarded for bravery, décoré pour le courage, *награжденный за храбрость*

* Здесь замкнутость понимается как внутренняя целостность и отграниченность объекта от других объектов. Так, например, английское in the arena указывает на нахождение в пределах арены, на арене, а не в каком-либо другом месте. Т. е. предлог in передает отношение к замкнутому пространству. В то же время on the arena (ср. СемП 7—100100) обозначает положение предмета на арене как на некоторой поверхности без указания на замкнутость пространства.

Таблица 4 (*продолжение*)

Код СемП'а	Название СемП'а	Примеры: английский, французский, русский
7—001	мотивация	rewarded for bravery, décoré pour le courage, *награжденный за храбрость*
7—000	немотивация	a building with columns, le bâtiment à colonnes, *здание с колоннами*
7—0011	причина	rewarded for bravery, décoré pour le courage, *награжденный за храбрость*
7—0010	непричина	fight for independence, lutter pour l'indépendance, *бороться за независимость*
7—00101	цель	fight for independence, lutter pour l'indépendence, *бороться за независимость*
7—00100	нецель	

7—000 немотивация

7—0001	совместность (т. е. общность, связанность)	a building with columns, le bâtiment à colonnes, *здание с колоннами*
7—0000	несовместность (т. е. разобщенность, несвязанность)	a message to the government, le message au gouvernement, *послание правительству*
7—00001	адресатность	a message to the government, le message au gouvernement, *послание правительству*
7—00000	неадресатность	write with: a pen, écrire avec le stylo, *писать ручкой*

Примечание. Для типов СемП'а, используемого относительно нульмерного, одномерного, двухмерного, трехмерного пространства и имеющего поэтому разные коды, приводится один общий лингвистический пример (ср. 7—111, 7—1011, 7—10011, 7—10001).

ние узлы тезауруса. ДСП, полученные путем анализа словарных дефиниций, синонимических рядов и контекста, например, «внутригосударственный институт — невнутригосударственный институт», «отношение к поверхности — отношение не к поверхности (к трехмерному пространству)», занимают нижние уровни иерархии (ср. табл. 1—4 и рис. 5—8).

Вместе с тем независимо от своей абстрактности каждый признак имеет двойную семиологическую значимость: признак является различительным (дифференцирующим) на одной ступени группировки лексических единиц и объединяющим (отождествляющим) на следующем за ним уровне абстракции.

Выявление отсутствующих звеньев тезауруса, а также проверка правильности и объективности выделенных СемП'ов и ДСП

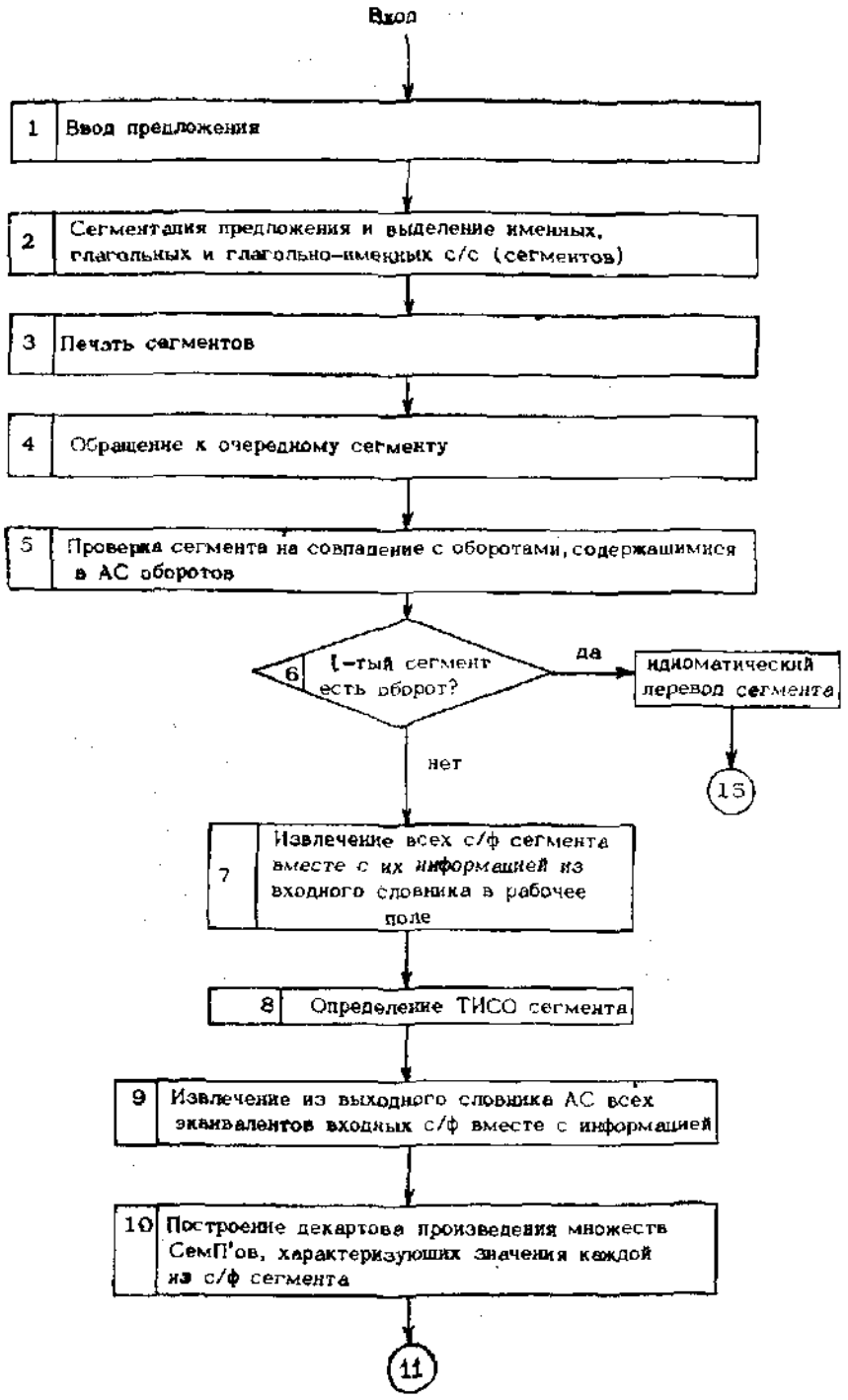


Рис. 9. Принципиальная блок-схема ФРС.

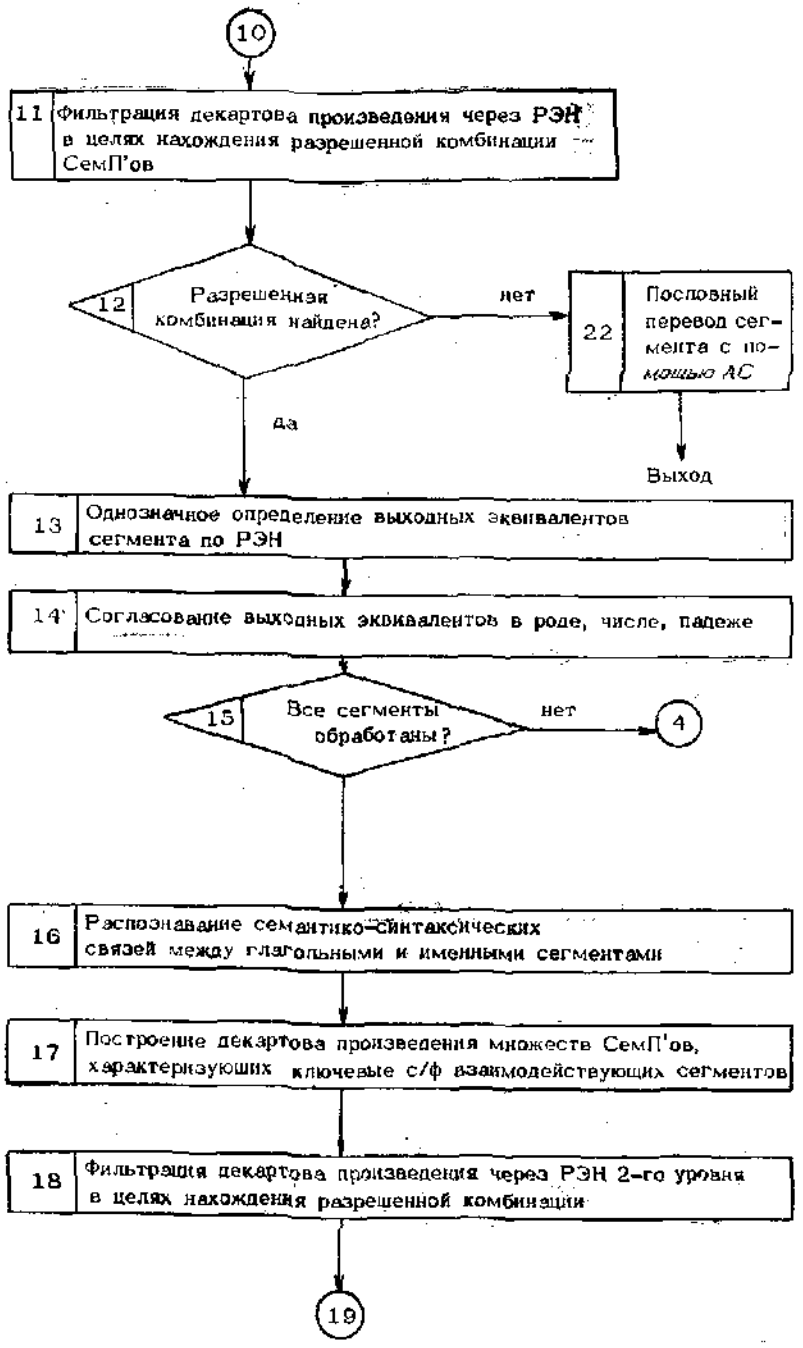


Рис. 9 (*продолжение*).

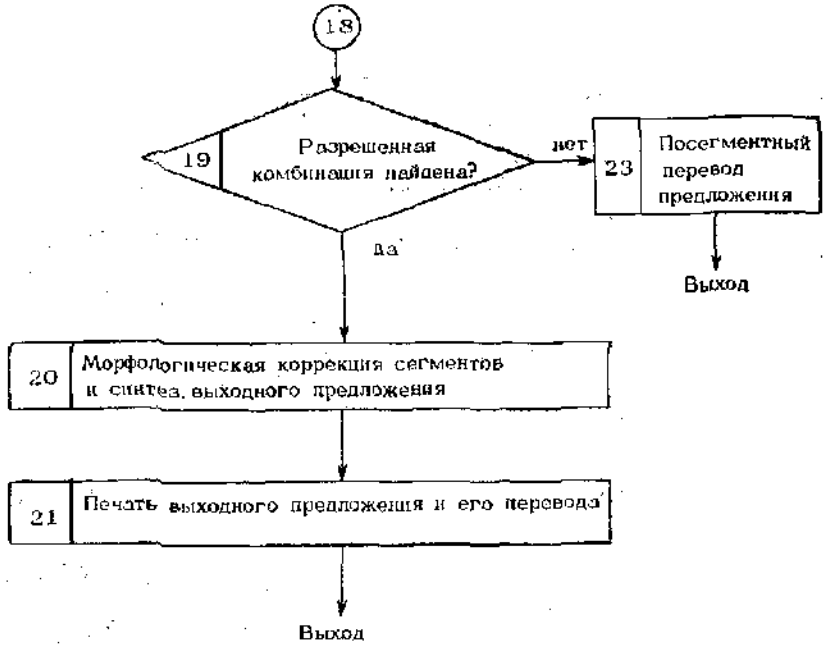


Рис. 9 (*продолжение*)

осуществляются исходя из тех энциклопедических знаний, которыми владеет алгоритмизатор. Именно на основании этих соображений СемП «внутригосударственный институт» расшифровывается в виде ДСП «орган государственного управления — неорган государственного управления».

Исходя из всего сказанного и используя перечисленные выше приемы выделения семантических признаков, построено четыре тезаурусных сети, описывающие СП существительных, глаголов, прилагательных и предложно-падежных отношений (табл. 1—4, рис. 5—8).

4.5 Лингвистические свойства тезауруса и правила семантического кодирования

Построить, в обозримый срок и применить в инженерно-лингвистической практике универсальный макротезаурус, способный описывать всю систему семантических объектов и связей, используемую в языках мира, пока невозможно. Поэтому приходится применять микротезаурусы, ориентированные на семантику определенных подъязыков и частей речи основных индоевропейских языков западной и восточной Европы. В данной работе, как уже

говорилось, используются четыре тезауруса, которые, будучи привязанными к основным частям речи (табл. 1—4, рис. 5—8), ориентированы на житейскую и общественно-политическую тематику. Как уже было сказано, на верхних уровнях каждого тезауруса располагаются наиболее абстрактные и универсальные ДСП и образующие их СемП'ы. По мере продвижения на нижние уровни тезауруса в него включаются идиоэтнические объекты, т. е. такие СемП'ы и ДСП, которые характерны для определенных языков и подъязыков. Как уже говорилось, глубина тезауруса (т. е. количество уровней ветвления) зависит от задач исследования и изученности семантики данной предметной области. Это значит, что тезаурус может, с одной стороны, развертывать новые уровни ветвления путем бинарного расщепления СемП'ов, находящихся на его терминальных узлах. С другой стороны, тезаурус может быть свернут путем отсечения всех ветвей, выходящих из некоторого фиксированного узла графа. Эти особенности обеспечивают адаптацию тезауруса не только к определенным семантическим исследованиям, но и к задачам кодирования конкретных лексических единиц.

Рассмотрим основные особенности и возможности тезауруса на примере механизма кодирования с/ф *франк*. Выработанный с помощью таблицы 5 код 2—0000000111, с одной стороны, отражает последовательность шагов кодирования с/ф *франк* по тезаурусу (ср. рис. 5), а с другой — указывает на семантическую соотнесенность этой словоформы с другими ЛЕ. Например, с/ф *вексель* получит согласно нашему тезаурусу код 2—0000000110, который указывает на то, что речь идет о денежном документе, но не денежной единице.

Код 2—0000000111, разумеется, не дает исчерпывающего описания семантики *франк*. Достаточно указать на то, что этим кодом будут обозначены также такие денежные единицы, как *доллар*, *су*, *экю*. Чтобы получить код, дающий полное описание семантики с/ф franc и однозначно выделяющий этот термин среди других обозначений денежных единиц, необходимо расширить этот код на основе дальнейшего ветвления тезауруса от терминального узла «денежная единица» путем введения таких ДСП, как «французская денежная единица — нефранцузская денежная единица», «современная денежная единица — старая денежная единица», «разменная (мелкая) монета — неразменная монета». В условиях нашей процедуры такая детализация кода является излишней, поскольку распознавание точного значения иноязычных эквивалентов слова *франк* осуществляется через АС.

С другой стороны, свертывание рассматриваемого куста тезауруса и устранение ДСП «денежная единица — неденежная единица», «денежный документ — неденежный документ» также недопустимо в нашем случае, поскольку это приводит к потере информации, необходимой для распознавания смысла лексической единицы.

Таблица 5

Последовательность кодирования слова *франк* по тезаурусу «Имя существительное»

№ кодовой позиции, соответствующей уровню тезауруса	Название лексико-грамматического класса или СемП'а	Код
1	существительное	2— *
2	«неконкретное»	0
3	«непроцесс»	0
4	«непризнак»	0
5	«несостояние»	0
6	«неотношение»	0
7	«неинститут»	0
8	«неотрасль народного хозяйства»	0
9	«финансово-экономическая категория»	1
10	«денежный документ»	1
11	«денежная единица»	1

Код с/ф *франк*: 2—0000000111

* Все начальные тезаурусные коды, совпадающие с принятыми в группе «Статистика речи» грамматическими кодами соответствующих частей речи (Пиотровский, 1975, с. 289, табл. 67, стлб. 1), даются в десятичной нотации, что отличает их от семантических кодов, представляемых в двоичном счислении.

§ 5. Построение распознающих фильтров

При описании математической модели ФРС (§ 3) указывалось, что выбор осмысленных словосочетаний в массе порождаемых автоматом многокомпонентных цепочек становится возможным только в том случае, если все эти цепочки пропускаются через лингвистический фильтр, который способен отделять осмысленные с/с от не имеющих смысла комбинаций с/ф. В качестве таких фильтров используются таблицы распознающих эталонных наборов (РЭН) разрешенных комбинаций кодов семантических признаков. Эти таблицы составляются на основе предварительного обследования сочетаемости прилагательных и существительных ($A+N \sim a+n$), двух существительных ($N_1+N_2 \sim n_2+n_1$ — см. табл. 6, также Билан, 1975, с. 90—96), глаголов и существительных ($V+N \sim v+n$ — см. табл. 7), глаголов, предлогов и существительных ($V+Prep+N \sim v+prep+n$ — см. табл. 8, вкл.).

Построение именных групп ($A+N$, $N_1+N_2 \sim a+n$, n_2+de+n_1; $A+N_1+N_2 \sim a+n_2+de+n_1$; и т. д.) синтаксически неизоморфно

в различных языках и имеет исключительно разветвленную и сложную семантическую сочетаемость. Поэтому целесообразно сначала отнести каждую именную группу к определенному типу интеркомпонентных смысловых отношений (ТИСО) — см. табл. 9. После этого именное с/с передается для распознавания его значения на той таблице РЭН, которая соответствует выявленному ТИСО (ср. табл. 6).

Таблица 6

Фрагмент распознающего эталонного набора (РЭН) разрешенных комбинаций СемП'ов для словосочетаний, передающих уточняюще-определительные отношения (D) и построенных по схеме «существительное + существительное» [англ. franc stability ($N_1 + N_2$), русск. *устойчивость франка* ($n_2 + n_1$)] и «прилагательное + + существительное» [англ. currency reform, русск. *валютная реформа* ($N_1 + N_2 \rightarrow a + n_2$)]

СемП'ы имен существительных ($N_1 = n_2$) и прилагательных ($N_1 = a$)	СемП'ы имен существительных $N_2 = n_1$: отдельное лицо	спецпроцесс	неспецпроцесс	неторгово-промышленная операция	свойство	равенство	орган управления
1	2	3	4	5	6	7	8
неотдельное лицо (группа лиц)					+		+
топоним							+
неспецпроцесс	+						
торгово-промышленная операция	+						
свойство							
неравенство							+
временной параметр							
состояние	+						+
отношение				+			+
орган управления	+			+			+
отрасль народного хозяйства	+				+		+
финансово-экономическая категория	+						+
денежная единица	+				+	+	
форма информации	+						
относящийся к финансово-экономической категории		+					

Для выявления ТИСО каждая именная форма снабжается во входном словнике АС записью, указывающей на то, в каких ТИСО может участвовать данная с/ф, принимая во внимание позицию, которую она занимает в словосочетании. Схема этой записи показана в таблице 10.

Таблица 7

Фрагмент распознающего эталонного набора разрешенных комбинаций СемП'ов в с/с «глагол + существительное»

СемП'ы глаголов	СемП'ы имен существительных: дискретное	недискретное	артефакт	отдельное лицо	неотдельное лицо	место	административно-географический объект	неадминистративно-географический объект	процесс	свойство	состояние	отношение	институт	предприятие	орган управления	торгово-промышленная операция	финансово-экономическая категория	денежный документ	отрасль народного хозяйства	форма информации
перемещение объекта	+	+	+	+	+									+				+		+
передача о. лицу (объекту)	+	+	+	+	+		+							+	+		+	+		+
производство материального о.	+	+	+											+				+		+
конструктивность идеального о.									+	+	+	+	+		+	+				
генерация о.							+		+	+	+	+	+							
модификация о.	+	+	+			+	+	+	+	+	+	+	+	+	+	+		+	+	+
увеличение о.	+					+	+	+	+	+	+	+		+		+		+	+	+
консолидация о.						+	+		+	+	+	+			+		+		+	+
уменьшение о.	+					+	+	+	+	+	+	+		+		+		+	+	+
физическое воздействие	+	+	+			+												+		
нефизическое в.				+					+	+	+	+	+				+		+	+
уничтожение объекта	+		+	+	+	+	+							+				+		+
элиминация									+	+	+	+	+	+	+	+	+	+	+	+
противодействие объекта				+	+		+		+	+	+	+	+	+	+	+		+		
непротиводействие (содействие) о.				+	+				+	+	+	+	+		+	+		+	+	+
агрессивное воздействие				+	+	+	+	+	+	+	+	+	+	+	+	+				
управление объектом					+	+	+		+				+	+	+	+			+	+
директивность				+	+		+		+					+	+				+	+
утилизация	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
объединение	+				+	+	+	+					+	+	+				+	+
необъединение (разъединение)	+				+	+	+	+					+	+	+				+	+
экспозитивность	+	+	+			+			+	+	+	+		+		+	+	+	+	+
темпоральность									+			+				+		+		
квалификативность	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
неквалификативность (квантитативность) →	+			+										+			+	+		+

Таблица 9

Типы интеркомпонентных смысловых отношений в именных группах

№ п/п	Название отношений	Код	Примеры: английский, французский, русский
1	уточняюще-определительные	D	defence Minister, ministre de la défense nationale, *министр обороны*
2	принадлежности	P	London area, territoire de Londres, *территория Лондона*
3	субъектные	S	government statement, déclaration du gouvernement, *заявление правительства*
4	объектные	O	leaflet distribution, distribution des tracts, *распространение листовок*
5	временные	T	May events, événements du mai, *майские события*
6	пространственные	L	Moscow congress, congrès de Moscou, *московский съезд*
7	оценочные	Ql	grave price situation, une grave situation des prix, *серьезное положение цен*
8	количественные	Qn	numerous country problems, grand nombre de problèmes du pays, *многочисленные проблемы страны*

Таблица 10

Участие английских с/ф franc и stability в различных ТИСО в зависимости от их позиций в именной группе $N_1 + N_2$ (0 — неучастие с/ф в заглавном ТИСО, 1 — участие с/ф в данном ТИСО)

С/ф	Позиции видов ТИСО в именном с/с															
	N_1								N_2							
	D	P	S	O	T	L	Ql	Qn	D	P	S	O	T	L	Ql	Qn
franc	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0	1
stability	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0

§ 6. Алгоритм формального распознавания смысла в системе МП

Принципиальная блок-схема МП, включающая формальное распознавание смысла входных с/у и с/с, показана на рис. 9. Ее блоки, обеспечивающие сегментацию входного текста, поиск по АС, морфологический и синтаксический синтез, много раз описывались в научной литературе (Вертель и др., 1971; Пиотровский, 1975, с. 83—91; Автоматическая переработка. . ., 1978; Пиотровский, 1979, с. 25—30, 86—87) и не нуждаются в новых коммента-

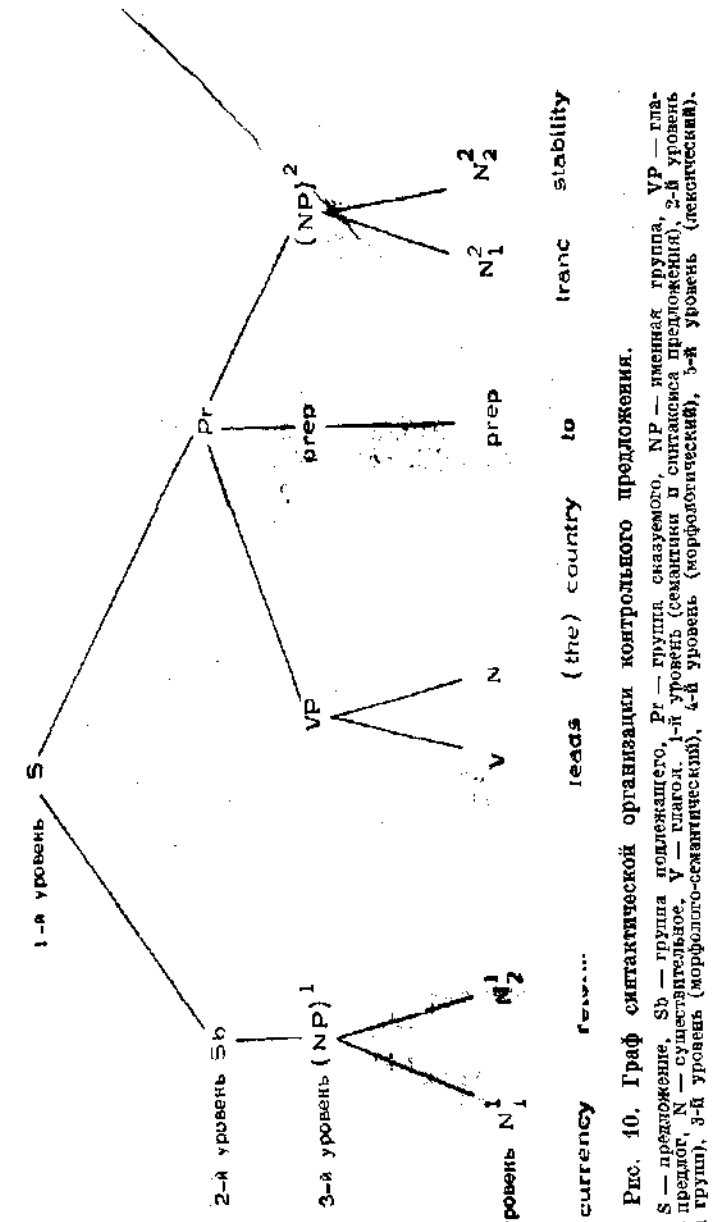


Рис. 10. Граф синтактической организации контрольного предложения.

Условные обозначения: S — предложение, Sb — группа подлежащего, Pr — группа сказуемого, NP — именная группа, VP — глагольная группа, prep — предлог, N — существительное, V — глагол. 1-й уровень (семантики и синтаксиса предложения), 2-й уровень (семантики и синтаксиса групп), 3-й уровень (морфолого-семантический), 4-й уровень (морфологический), 5-й уровень (лексический).

риях. Поэтому мы сосредоточим внимание на описании блоков 7—20, в которых реализуется математическая модель $2|^{(2)}$ (§ 3), обеспечивающая распознавание смысла входных с/у и с/с средствами выходного языка путем анализа сочетаемости семантических кодов на морфолого-семантическом уровне и на уровне семантики и синтаксиса групп подлежащего и сказуемого.

Работу этих блоков проследим на примере перевода английского предложения The currency reform leads the country to franc stability, синтаксическая организация которого показана на рис. 10.

Сегментация, осуществляемая блоком 2, разбивает предложение на простые именные и глагольные сегменты, занимающие 3-й слой синтаксического графа. Если сегмент не является машинным оборотом (ср. блоки 6, 19, 20), то он поступает в блоки 7—16, реализующие алгоритм ФРС, который действует в одном из двух режимов в зависимости от того, обрабатывается именной или глагольный сегмент.

6.1 Распознавание смысла с/ф в именной группе $N_1+N_2(A+N)$

В контрольном предложении имеется две именных группы, построенных по заглавной схеме — currency reform и franc stability. Их обработка начинается с определения присущих им типов интеркомпонентных смысловых отношений. Для этого кодовые записи обеих с/ф в каждом словосочетании сравниваются на совпадение присущих им ТИСО (табл. 11). Рассмотрим механизм этой

Таблица 11

Определение типа интеркомпонентного смыслового отношения в сегментах currency reform и franc stability

Сегменты	D	P	S	O	T	L	Q1	Qn
N_1 currency	1	0	0	1	0	0	0	0
N_2 reform	1	0	0	0	1	1	1	1
N_1 franc	1	0	0	1	0	0	0	0
N_2 stability	1	0	0	0	0	0	1	0

процедуры на примере с/с franc stability. С/ф franc занимает первое место в сегменте. Поэтому в таблице 11 перенесена его кодовая характеристика из столбца N_1 таблицы 10. С/ф stability получает соответственно запись из столбца N_2 той же таблицы. Оба существительных имеют одновременно единицы только в столбце D, поэтому образованное из них с/с franc stability характеризуется

уточняюще-определительным отношением (D). Этим же способом выясняется, что с/с currency reform также содержит то же отношение D.

Таблица 12

Переводные эквиваленты и СемП'ы английских словоформ currency, reform, franc, stability

№№ сегментов	Входные с/ф	Переводные эквиваленты	Название СемП'ов	Коды СемП'ов
1	currency	*валюта, деньги*	финансово-экономическая категория	2—00000001
		распространенность	свойство	2—0011
		продолжительность	временной параметр	2—0010101
		валютный, денежный	относящийся к финансово-экономической категории	3—01000001
	reform	*реформа*	спецпроцесс	2—011
		исправление	неспецпроцесс	2—010
2	franc	*франк*	денежная единица	2—0000000111
	stability	*устойчивость*	свойство	2—0011

Теперь обратимся к рассмотрению формальных пар СемП'ов, характеризующих оба словосочетания. По данным таблицы 12 видно, что с/с currency reform дает $4\times2=8$ пар, а franc stability только одну пару СемП'ов.

В результате сравнения первых восьми пар с D-РЭН (табл. 6) выясняется, что разрешенной, т. е. зафиксированной в D-РЭН парой является комбинация адъективного и именного СемП'ов («относящийся к финансово-экономической категории», «спецпроцесс»), определяющая те русские ЛЕ, которые в данной структуре соответствуют английским currency и reform. Ими являются словоформы *валютный* (*денежный*) и *реформа*, которые, получив в блоке 14 нужное синтаксико-морфологическое оформление, сформируют два варианта семантически правильного перевода: *валютная реформа*, *денежная реформа*. Поскольку алгоритм пока не способен производить стилистический выбор между двумя синонимическими ЛЕ, то автомат выдает на печать либо оба варианта, либо только первый вариант перевода.

Сегмент franc stability характеризуется одной парой именных СемП'ов («денежная единица», «свойство»), засвидетельственной

в D-РЭН. Это приводит ЛА к однозначному выбору русской инверсивной схемы n_1+n_2, которая в блоке 14 получает правильное синтаксико-морфологическое оформление *устойчивость франка*.

6.2. Распознавание смысла с/ф в глагольно-именной группе V+N

Распознавание смысла в группе V+N осуществляется по более простой схеме, чем распознавание в именном с/с. Интеркомпонентные смысловые отношения здесь не определяются, а все пары СемП'ов, входящие в декартово произведение V×N, фильтруются через единый глагольно-именной РЭН (табл. 7).

В результате такой фильтрации через РЭН восьми пар СемП'ов, полученных из декартова произведения множеств leads×country (см. табл. 13), выясняется, что две пары («управление», «административно-географический объект» и «директивность», «административно-географический объект») оказываются отмеченными в РЭН. Это значит, что входной сегмент leads the country to имеет, учитывая работу блока 14, два переводных эквивалента — *руководит страной*, *ведет страну*.

Таблица 13

Переводные эквиваленты и СемП'ы английский с/ф leads и country

Входные с/ф	Переводные эквиваленты	Название СемП'ов	Коды СемП'ов
leads	*руководит*	управление объектом	4—1100001
	ведет	директивность	4—1100011
	убеждает	нефизическое воздействие	4—1100110
country	*страна*	административно-географический объект	2—11011
		неадминистративно-географический объект	2—11010

Устранить эту неоднозначность входного сегмента на данном уровне анализа не удается. Ее можно снять лишь на более высоком уровне семантико-синтаксического анализа, учитывающего предложное управление глагола to lead.

6.3. Распознавание семантико-синтаксических связей между глагольными и именными сегментами в предложении

В предыдущих разделах было показано, что распознавание смысла ЛЕ, в рамках того сегмента, к которому принадлежит данная ЛЕ, не всегда приводит к однозначному и правильному решению. Поэтому результаты, полученные на морфолого-семантическом (третьем) уровне анализа, должны проверяться и кор-

Таблица 8

Фрагмент распознающего эталонного набора разрешенных комбинаций СЕМП'ов в с/с «глагол + предлог + существительное»

Глагольный СемП	Именной СемП															
	артефакт	одушевленность	место	процесс	свойство	количественный параметр	временной параметр	состояние	институт	предприятие	непредприятие (учреждение)	отрасль народного хозяйства	форма информации	форма мыслительной деятельности	денежная единица
перемещение объекта		начальная граничность, неначальная (конечная) гр.	начальная граничность, неначальная (конечная) гр., преодоление пространства							начальная граничность, неначальная (конечная) гр.					
передача о. лицу		адресатность							адресатность						
неотчуждение о. (уступительность)		адресатность							адресатность						
производство материального о.	инструментальность		нахождение в пределах пространства, нахождение вне пр.	цель						нахождение в пределах пространства					
конструктивность идеального о.			нахождение в пределах пространства, нахождение вне пр.	цель					нахождение в п	делах пространства					
увеличение о.						начальная граничность, предельность, квантитативность, преодоление пр.									начальная граничность, предельность (квантитативность)
уменьшение о.						начальная граничность, предельность, квантитативность, преодоление пр.									начальная граничность, предельность (квантитативность)
физическое воздействие	инструментальность	конечная граничность													
объединение объектов		совместность								совместность					
экспозитивность о.		адресатность							адресатность						
управление о.		совместность		цель					совместность				соответствие, противоречие		
директивность		конечная граничность			непредельность относительно нульмерного пространства (точки)			конечная граничность							
уменьшение о.						начальная граничность, предельность, квантитативность, преодоление пр.									начальная граничность, предельность (квантитативность)
физическое воздействие	инструментальность	конечная граничность													
объединение объектов		совместность								совместность					
экспозитивность о.		адресатность								адресатность					
управление о.		совместность		цель						совместность			соответствие, противоречие		
директивность		конечная граничность			непредельность относительно нульмерного пространства (точки)					конечная граничность					
санкция				причина									соответствие, противоречие		
торгово-финансовая акция	компенсационность														компенсационность
утилизация объекта			нахождение в пределах пространства, нахождение вне пр.	цель						нахождение в пределах пространства					
мыслительная деятельность	указание на содержание														
информативная деятельность		совместность, адресатность	начальная граничность, неначальная (конечная) гр.							совместность, адресатность					
	указание на содержание														
состояние субъекта		совместность, несовместность	нахождение в пределах пространства, нахождение вне пр.						нахожден	не в пределах пространства					
									темпоральность	причина, совместность, несовместность					
процесс									темпоральность	причина					
функциональность субъекта		совместность, несовместность	нахождение в пределах пространства, нахождение вне пр.	цель	совместность, несовместность				темпоральность	совместность, несовместность	нахождение в пределах пространства		соответствие, противоречие		

Примечание. В клетках на пересечении глагольных и именных СемП'ов находятся СемП'ы предлогов. Помимо семантических признаков, показанных на фрагменте предложного дерева, в таблице сочетаемости фигурируют также некоторые предложные СемП'ы, не указанные на рис. 8.

ректироваться на более высоком (втором) уровне распознавания — уровне семантики и синтаксиса групп.

Рассмотрим эту задачу на примере распознавания отношений между сегментами leads the country to — *руководит страной, ведет страну*, franc stability — *устойчивость франка*.

Задача состоит в том, чтобы отыскать с помощью РЭН такую цепочку СемП'ов, которая соответствовала бы тем семантико-синтаксическим отношениям, которые существуют между ключевыми ЛЕ указанных сегментов. Иными словами, эта цепочка является средством распознавания связи, существующей в данном контексте между с/ф leads, to, (the), stability. Исходными данными для решения указанной задачи являются:

перевод существительного stability и соответствующий СемП вместе с его кодом;

переводы глагола leads и соответствующие им СемП'ы и коды;

множество переводных эквивалентов предлога to вместе с их СемП'амн и кодами.

На основании этих данных, приведенных в таблице 14, строится декартово произведение leads×to×(the)×stability, дающее 14 (2·7·1) троек СемП'ов. Фильтрация этих троек через распознаю-

Таблица 14

Переводные эквиваленты и СемП'ы английских с/ф leads, to, (the), stability

Входные с/ф	Переводные эквиваленты	Название СемП'ов	Коды СемП'ов
leads	*руководит*	управление объектом	4—1100001
	ведет	директивность	4—11000011
to	*к*	непредельность по отношению к нульмерному пространству (точке)	7—11010100
	на + вин. пад.	непредельность по отношению к двухмерному пространству	7—10010010100
	в + вин. пад.	непредельность по отношению к трехмерному пространству	7—10000110100
	до	предельность по отношению к нульмерному пространству (точке)	7—11010101
	для	цель	7—00101
	с + тв. пад.	совместность	7—0001
	флексия дательного пад. зависимого существительного	адресатность	7—00001
(the) stability	*устойчивость*	свойство	2—0011

щий эталонный набор, приведенный в таблице 8, обнаруживает одну отмеченную в системе и норме английского и русского языков тройку («директивность», «непредельность относительно одно-

```
THE CURRENCY REFORM LEADS THE COUNTRY TO FRANC STABILITY
CURRENCY REFORM
1 ЭТАП: ОПРЕДЕЛЕНИЕ ТИСО
УСТАНОВЛЕН D-ТИСО
2 ЭТАП: ОПРЕДЕЛЕНИЕ ФОРМУЛЫ ПЕРЕВОДА
СОЧЕТАНИЕ ПЕРЕВОДИТСЯ ПО ФОРМУЛЕ A1+N2
PR2 = РЕФОРМА
PR1 = ВАЛЮТНАЯ
PR1 = ДЕНЕЖНАЯ
CURRENCY REFORM - ВАЛЮТНАЯ = ДЕНЕЖНАЯ РЕФОРМА


FRANC STABILITY
1 ЭТАП: ОПРЕДЕЛЕНИЕ ТИСО
УСТАНОВЛЕН D-ТИСО
2 ЭТАП: ОПРЕДЕЛЕНИЕ ФОРМУЛЫ ПЕРЕВОДА
СОЧЕТАНИЕ ПЕРЕВОДИТСЯ ПО ФОРМУЛЕ N2+N1
PR2 = ФРАНКА
PR1 = УСТОЙЧИВОСТЬ
FRANC STABILITY - УСТОЙЧИВОСТЬ ФРАНКА


LEADS THE COUNTRY
ОПРЕДЕЛЕНИЕ НОМЕРА РАСПОЗНАВАЕМОЙ МАТРИЦЫ
МАТРИЦА N1 N2
PR1 = РУКОВОДИТ
PR1 = ВЕДЕТ
PR2 = СТРАНА
LEADS THE COUNTRY - РУКОВОДИТ СТРАНОЙ = ВЕДЕТ СТРАНУ


LEADS TO STABILITY
ВЕДЕТ = РУКОВОДИТ ТО УСТОЙЧИВОСТЬ
PR1 = ВЕДЕТ
PR2 = К
PR3 = УСТОЙЧИВОСТИ
LEADS TO STABILITY - ВЕДЕТ К УСТОЙЧИВОСТИ


LEADS THE COUNTRY TO FRANC STABILITY
ВЕДЕТ СТРАНУ К УСТОЙЧИВОСТИ


THE CURRENCY REFORM LEADS THE COUNTRY TO FRANC STABILITY
ВАЛЮТНАЯ + ДЕНЕЖНАЯ РЕФОРМА ВЕДЕТ СТРАНУ К УСТОЙЧИВОСТИ ФРАНКА
```

Рис. 11. Семантический и морфолого-синтаксический МП английского предложения the currency reform leads the country to franc stability.

мерного пространства», «свойство»), которая указывает на то, что выходной эквивалент английского leads to (the) stability должен включать с/ф *ведет*, *к*, *устойчивость*.

6.4. Работа блока морфологической коррекции сегментов и синтеза выходного предложения

По контрольному предложению здесь производят две операции:

1. Синтаксико-морфологическое упорядочение полученного сегмента *ведет к устойчивости* с включением в него с/ф *страна* и *франк*. В результате этой операции два входных сегмента приобретают вид: *ведет страну к устойчивости франка*.

2. Морфолого-синтаксическое объединение перевода сегмента the currency reform с переводом сегмента leads the country to franc stability.

В результате формируется полный МП предложения (рис. 11).

§ 7. Заключение

Мы построили процедуру формального распознавания смысла словоформ и словосочетаний текста, включающую математическую модель, лингвистическое обеспечение, алгоритм и его программную реализацию на ЕС ЭВМ. Эта процедура представляет собой реализацию второго — семантико-морфологического и отчасти третьего — семантико-синтаксического уровней, надстраивающихся над уже реализованным первым — лексическим уровнем ЛА.

Построение многоуровневого лингвистического автомата представляет собой не только с технологической, но и с концептуальной точки зрения трудную и внутренне парадоксальную задачу. Парадоксальность этой задачи состоит уже в том, что наши знания о построении языка и механизмах построения текста относятся в основном к низшим лингвистическим уровням (в первую очередь к лексическому), что же касается верхних уровней лингвистической архитектуры — ситуативному, коммуникативному, то здесь вместо твердых и надежных знаний мы располагаем лишь зыбкими гипотезами. Между тем, в глобальном ЛА модели верхних уровней должны управлять моделями низших уровней.

Перед лицом этого онтолого-гносеологического парадокса мы должны либо прекратить практические работы по построению глобального лингвистического автомата, ожидая накопления достаточного фактического материала и теоретических разработок, либо продолжать его построение, начиная с низших его уровней и помня при этом, что модели этих уровней будут в дальнейшем корректироваться и перестраиваться по мере появления воспроизводящих моделей более высоких уровней.

Социальные требования к инженерной лингвистике, информатике и проблематике искусственного разума таковы, что единственным правильным путем является путь последовательного развития ЛА от низших уровней к высшим, хотя этот путь связан с большими ментальными, техническими и организационно-финансовыми затратами.

Часть II
СТАТИСТИКА РЕЧИ

С. А. Шубик

СТАТИСТИЧЕСКИЕ МЕТОДЫ В ЛИНГВИСТИКЕ

Одна из характерных особенностей современной лингвистики — широкое использование статистических методов. Об этом хорошо сказал известный болгарский лингвист Н. В. Ставровский: «Если бы мне предложили сразу, без подготовки, перечислить то, что характеризует современное языкознание, наверное, мое перечисление оказалось бы неполным, я, наверное, опустил бы одну или другую тенденцию, типичную для языковедов нашего времени — настолько современное состояние языкознания сложно, многогранно, но, конечно, не опустил бы в качестве его характерной черты интереса к применению статистического метода» (Ставровский, 1971, с. 80). Большой интерес современного языкознания к статистике обусловлен в основном следующими обстоятельствами: 1) более глубоким пониманием многих явлений языка и речи; 2) бурным развитием лингвистических дисциплин, которые испытывают повышенную потребность в статистических методах (функциональная стилистика, социолингвистика, психолингвистика, лингвистическая география); 3) появлением прикладных задач, связанных с автоматической переработкой текста (автоматический перевод, автоматическое реферирование и др.); 4) усиленными поисками новых путей увеличения объективности и точности исследования.

Современное языкознание характеризуется, однако, не только широким использованием статистики, но и широким обсуждением статистической методологии. Немало работ — не только статей, но и книг — посвящено применению в лингвистике вероятностно-статистических методов (математическая статистика) (Guiraud, 1960; Herdan, 1966; Савицкий, 1966; Зиндер, Строева, 1968; Адмони, 1970; Головин, 1971; Фрумкина, 1975; Пиотровский, Бектаев, Пиотровская, 1977; и др.). Тем не менее остается еще во многих отношениях неясным вопрос о границах их приложимости. Этот вопрос — одна из главных тем данной статьи.

Другая главная тема — специфика качественно-количественных методов и их роль в лингвистических исследованиях. Качественно-количественные методы нередко отвергаются или игнорируются, хотя теория и практика говорят о том, что они необходимы для решения разнообразных лингвистических задач.[1]

Мы коснемся также вопроса о сотрудничестве вероятностно-статистических и качественно-количественных методов.

Необходимой предпосылкой применения математической статистики является образование репрезентативной выборки на основе принципа случайности, в соответствии с которым всем единицам генеральной совокупности должна быть предоставлена одинаковая возможность попасть в выборку. Выполнимость принципа случайности связана с характером генеральной совокупности, с ее качеством и объемом.

По качеству генеральные совокупности делятся на статистически однородные (однокачественные) и статистически неоднородные (разнокачественные). В однородной совокупности на разных ее участках действует один и тот же комплекс систематических факторов (или один и тот же систематический фактор). Именно это постоянство комплекса систематических факторов обеспечивает близость статистических показателей различных выборок, взятых из одной и той же генеральной совокупности. Что же касается количественных расхождений между различными выборками однородной совокупности, то они создаются так называемыми случайными факторами, т. е. такими, которые хаотически то усиливают, то ослабляют действие постоянного комплекса систематических факторов. В неоднородной совокупности комплекс систематических факторов непостоянен, на разных ее участках он так или иначе изменяется: либо какие-то факторы прекращают свое действие, либо появляются новые факторы, либо происходит и то и другое. Статистические показатели различных выборок неоднородной совокупности значительно отличаются друг от друга. Математическая статистика располагает специальными средствами, предназначенными для отграничения однородного распределения выборочных показателей от неоднородного.

По объему генеральные совокупности делятся на конечные и бесконечные. «Если совокупность содержит ограниченное число единиц, она называется конечной. Если число единиц совокупности безгранично, ее называют бесконечной совокупностью» (Венецкий, Кильдишев, 1975, с. 215). В лингвистике целесообразно придать этим терминам несколько иной смысл. Конечной совокупностью мы будем считать практически обозримый отрезок речи, относительно небольшой, четко очерчен-

[1] См.: Адмони, 1964, с. 62—72, 78—83; Ulvestad, 1962. Из старых работ особого внимания заслуживают статьи Д. Н. Кудрявского (Кудрявский, 1909, 1911, 1912).

ный, целиком доступный для статистической обработки (отдельная книга, ограниченный список сочинений одного или разных авторов, сумма статей на определенную тему из определенного журнала за определенный срок и т. п.). Бесконечной совокупностью мы будем называть практически необозримые отрезки речи — все тексты данного языка или его какой-нибудь функционально-стилистической разновидности не только вне хронологических рамок (в этом случае мы имеем дело с подлинно бесконечными совокупностями, ибо тексты языка и его функционально-стилистических разновидностей постоянно пополняются), но и в определенный, даже сравнительно небольшой период времени (например, за три года).

Какова же связь между характером генеральной совокупности и приложимостью вероятностно-статистических методов к ее изучению?

Если генеральная совокупность конечна, то математическую статистику можно использовать без каких-либо серьезных ограничений, потому что в этом случае существует, как правило, реальная возможность сделать достаточно репрезентативную выборку независимо от того, однородна генеральная совокупность или неоднородна. Наиболее уместным способом образования выборки в языкознании, как, впрочем, и в других общественных науках, является механический отбор, при котором вся генеральная совокупность разбивается на равные части и из каждой части берется либо одна единица (индивидуальный механический отбор), либо целая серия (серийный механический отбор). Механический отбор легко организовать, и он хорошо улавливает изменения в качестве генеральной совокупности. Если известно или есть основание полагать, что генеральная совокупность однородна или не слишком разнородна, то предпочтения заслуживает серийный отбор, поскольку он требует меньшей затраты труда и обычно облегчает качественный анализ. Но если генеральная совокупность очень неоднородна, серийный отбор может дать о ней ложное представление, здесь предпочтительнее индивидуальный отбор.

Выборочное обследование, в отличие от сплошного, никогда не дает абсолютно достоверного (абсолютно точного и абсолютно надежного) знания: характеристики генеральной совокупности, исчисленные на основе выборки, обязательно сопровождаются указанием на возможную величину ошибки репрезентативности и на определенную степень вероятности того, что эта ошибка не превысит данной величины. Математическая статистика позволяет рассчитать объем выборки, необходимый для обеспечения очень высокого, практически вполне достаточного уровня достоверности. Несущественный проигрыш в достоверности выборочного обследования окупается существенным выигрышем в производительности труда: объем работы сокращается во много раз (обычно выборка составляет от 5 до 20% генеральной совокупности). К тому же надо иметь в виду, что при очень большом объеме

генеральной совокупности качественный анализ всех ее единиц часто практически неосуществим или во всяком случае не может быть проведен с надлежащей тщательностью. Сплошное обследование конечной генеральной совокупности оправдано только тогда, когда генеральная совокупность либо достаточно мала, либо настолько разнородна, что трудно гарантировать репрезентативность выборки.

Обратимся теперь к применению вероятностно-статистических методов при изучении бесконечной генеральной совокупности. Здесь их возможности существенно ограничены. В частности, для исчисления обобщающих характеристик бесконечной совокупности вероятностно-статистические методы приложимы только тогда, когда бесконечная совокупность однородна (в противном случае они способны лишь зафиксировать самый факт неоднородности). Связано это с тем, что бесконечная совокупность не может быть положена вся целиком в основу выборки; поэтому выборка производится из какого-нибудь фрагмента бесконечной совокупности, который, как бы ни был велик, составляет незначительную часть всей совокупности. Так, например, если исследователь (или исследовательский коллектив) преследует цель изучить научный стиль (определенного языка в определенную эпоху), он не может произвести отбор из всех книг и журналов всех научных дисциплин и вынужден ограничиться сравнительно небольшим списком текстов общим объемом в несколько сот или тысяч страниц. Но выборка, сделанная из какого-нибудь фрагмента бесконечной совокупности, непосредственно репрезентирует только данный фрагмент, и исчисленные на ее основе характеристики непосредственно относятся только к нему. Распространение характеристик фрагмента бесконечной совокупности на всю совокупность правомерно лишь в том случае, если качество фрагмента бесконечной совокупности совпадает с качеством всей совокупности, другими словами — если бесконечная совокупность однородна, так что ее способен замещать некоторый, взятый из нее наудачу фрагмент.

Как же обстоит дело с качеством бесконечных совокупностей языковых единиц? Рассмотрим этот вопрос на материале функциональных стилей современных европейских языков.[2]

Результаты многочисленных исследований (Савицкий, 1966; Статистичні параметри стилів, 1967; Журавлев, 1967; Лесскис, 1968; Кауфман, 1970; Кожина, 1972; Русова, 1974; Грехнева, 1974; Никонов, 1978; и др.) позволяют сделать вывод, что язык как совокупность функциональных стилей не имеет ни одного

[2] Мы не останавливаемся на спорах о количестве и составе функциональных стилей и подстилей. Существующие здесь неясности создают трудности в разграничении лингвостатистических совокупностей, определяемых через понятия функциональных стилей и подстилей. Чаще всего выделяют пять функциональных стилей: художественный, публицистический, научный, официально-деловой, обиходно-разговорный (ср.: Троянская, 1979, с. 4).

количественного признака, относительно которого он был бы однороден. Каждый функциональный стиль, как правило, также неоднороден. «Опыты показывают, — констатирует Б. Н. Головин, — что главные функциональные стили имеют внутреннюю дифференциацию на жанровые, тематические и даже индивидуальные разновидности. Так, разновидность публицистического стиля, соотнесенная с жанром заметок и сообщений, может по ряду признаков существенно отличаться от разновидности, соотнесенной с жанром передовой статьи: вероятности какого-то круга грамматических категорий и явлений лексики окажутся в газетных передовицах и в газетных информационных сообщениях дифференцирующими, т. е. существенно различными. Разновидность научного стиля, обслуживающая нужды теоретической физики, может оказаться отделенной некоторым набором дифференцирующих вероятностей (долей) от разновидности, обслуживающей нужды математики или биологии. Индивидуальная разновидность художественного стиля, формируемая и применяемая в прозе Л. Леонова, может существенно отличаться от индивидуальной разновидности того же стиля, формируемой и применяемой в прозе К. Паустовского» (Головин, 1971, с. 127—128).

Однородность в том или ином отношении нередко обнаруживают отдельные разновидности функциональных стилей. Так, например, по данным Т. А. Якубайтис и Б. А. Стурите, в латышской художественной литературе стиль прозы, стиль драматических произведений и стиль поэзии однородны по пяти признакам — по количеству разных грамматических форм, по количеству глаголов и существительных, по количеству существительных в винительном и дательном падежах (Якубайтис, Стурите, 1974).[3] Допустима гипотеза Л. Е. Машкиной о том, что в немецких газетных текстах политической тематики нет статистически значимых различий в распределении определенных сочетаний слов (Машкина, 1969). Согласно Р. Ю. Кобрину и Л. А. Пекарской, русским текстам по объемной штамповке, публикуемым на страницах реферативного журнала «Технология машиностроения», свойственно устойчивое распределение терминов, состоящих из одного, двух или более слов (Кобрин, Пекарская, 1977, с. 274—275).

Нельзя, однако, упускать из виду, что и отдельные разновидности функциональных стилей однородны далеко не всегда. Группа исследователей под руководством В. И. Перебейнос показала, например, что в украинской прозе, драматургии и поэзии отчетливо выделяются авторские стили на разных уровнях языка (изучались следующие признаки: частота классов фонем — гласных, согласных, сонорных, звонких, глухих и мягких; частоты

[3] Интересно отметить, что один из этих признаков, а именно частота глаголов, оказался однородным для латышского художественного стиля в целом.

глагольных форм — настоящего, прошедшего и будущего времени, повелительного наклонения, инфинитива, причастия и деепричастия; частоты сочинительных и подчинительных союзов; частоты знаков препинания — точки, многоточия, восклицательного и вопросительного знаков; средняя длина предложения и распределение размеров предложения; средняя длина слова и распределение размеров слова) (Перебейнос, 1968). Как установил Б. Я. Слепак, в английских журналах по машиностроению имеются существенные расхождения в употребительности некоторых синтаксических конструкций между такими разновидностями текста, как техническая характеристика механизмов, описание их устройства, информация относительно их функционирования, описание технологических процессов, изложение результатов научных исследований (Слепак, 1978, с. 10—11).

Необходимо обратить внимание на то, что некоторые исследователи склонны принимать гипотезу об однородности той или иной лингвостатистической совокупности без достаточных оснований. Так, А. П. Журавлев утверждает, что научный стиль русского языка однороден по относительной частоте придаточных предложений изъяснительного типа (Журавлев, 1967, с. 105), однако обследованный им материал [4] не может считаться репрезентативным, потому что в нем представлены выборки из сочинений только двух наук (математики и лингвистики) и, кроме того, почти не отражена дифференциация текстов в зависимости от раздела науки, жанра, функционально-смыслового типа (рассуждение, описание, повествование). Как показал В. Г. Адмони (Адмони, 1970, с. 95—98), недостаточно разнообразный материал привлекла и Т. В. Строева для доказательства тезиса о том, что разные тексты научного стиля как в русском, так и в немецком языках не имеют статистически значимых различий по употребительности родительного падежа имени существительного (Строева, 1968, с. 8—11).

Работу по изучению качества бесконечных совокупностей языковых единиц следует интенсифицировать. Нужно проверить качество самых разных совокупностей по самым разным признакам. Здесь необходимо идти не только по пути той или иной дифференциации функциональных стилей. Целесообразно также варьировать объем серий выборки и их число: в общем, чем больше объем и число серий, тем выше степень однородности. Нельзя, далее, упускать из виду, что однородное распределение чаще имеют родовые признаки, чем видовые: так, отдельные фонемы обычно распределяются неоднородно, тогда как некоторые классы фонем (например, класс согласных или класс гласных) — однородно. Степень однородности может быть также улучшена заменой абсолютной частоты на долю в массе однопорядковых явлений.

[4] Список обследованных текстов приводится А. П. Журавлевым в другой работе (Журавлев, 1969, с. 223).

Желательно, чтобы подтверждение гипотезы об однородности одной и той же совокупности в отношении одного и того же признака было получено разными исследователями на разном материале. Еще важнее, чтобы эмпирическое обнаружение однородности сопровождалось теоретическим обоснованием — это в высшей степени создает уверенность в том, что полученные в опыте результаты не являются случайными. «Однородна ли данная статистическая масса — этого нельзя установить на основе только статистических исследований, — справедливо пишут польские статистики О. Ланге и А. Банасиньский. — Нужен качественный анализ, который проводится методами, применяемыми в областях науки, где изучаются данные явления: в экономике, медицине, социологии, психологии и т. д.» (Ланге, Банасиньский, 1971, с. 25—26).

Исследования по изучению качества бесконечных лингвостатистических совокупностей важны не только потому, что они выделяют те аспекты речевой деятельности, при описании которых можно и нужно широко использовать вероятностно-статистические методы. Такие исследования содействуют также более четкой и надежной дифференциации различных типов речи и дают известное представление о характере действующих в них факторов (поскольку, как мы видели выше, механизм образования однородной и неоднородной совокупностей принципиально различен).[5]

До тех пор пока гипотеза об однородности бесконечной совокупности экспериментально и теоретически не подтверждена, исчисление ее характеристик с помощью математической статистики не может быть оправдано. Между тем многие лингвисты пренебрегают этим условием, в силу чего полученные ими результаты не заслуживают никакого доверия и, следовательно, не представляют никакой ценности. Так, например, И. С. Вольская и В. С. Горевая использовали аппарат математической статистики для исчисления среднего размера предложения в английской художественной прозе. Поскольку, однако, эта функционально-стилистическая разновидность по распределению размеров предложения неоднородна, они получили совершенно разные результаты: согласно И. С. Вольской, средний размер предложения в английской художественной прозе находится в пределах от 19.6 до 22.4 слова (*Вольская, 1965, с. 72*), согласно В. С. Горевой — от 16.5 до 19.4 (Горевая, 1969, с. 79).

Итак, многие бесконечные совокупности языковых элементов статистически неоднородны. Это означает, что свойственные им статистические закономерности недостаточно строги, недостаточно последовательны для того, чтобы их можно было описывать с помощью вероятностно-статистического аппарата. При изучении нестрогих статистических закономерностей, которые в силу их непоследовательности обычно именуются тенденциями, приходится обращаться к качественно-количественным методам.

Но непосредственная и основная область использования качественно-количественных методов — это изучение качественной стороны статистических совокупностей, тех многочисленных факторов, которые управляют каждой единицей совокупности и которые, различным образом комбинируясь и взаимодействуя, формируют статистическую закономерность. Вероятностно-статистические методы, если только они вообще приложимы, играют здесь вспомогательную роль.

Качественная специфика статистических совокупностей состоит в том, что в них действует множество разнообразных факторов, существенных и несущественных, систематических и случайных, внутренних и внешних, при этом число и состав факторов, действующих на каждую единицу совокупности, часто не совпадают, частично или полностью. Поэтому один или несколько случаев здесь не показательны для всей массы в целом, здесь необходимо планомерное изучение достаточно большого количества единиц, чтобы выявить более или менее полный перечень представленных в данной совокупности факторов, установить возможности их комбинирования и взаимодействия и, наконец, определить меру участия каждого фактора.

Важнейшую роль в качественно-количественном исследовании играют статистические группировки. Одна из таких группировок представлена в таблице 1, где показано употребление наклонения в немецких изъяснительных предложениях в зависимости от так называемых индифферентных глаголов в формах претерита и плюсквамперфекта.[6]

Из таблицы 1 видно, что в изъяснительных предложениях при отсутствии союза daß конъюнктив употребляется примерно в 7 раз чаще, чем при его наличии. А это означает, что наличие/отсутствие союза daß является важным фактором выбора наклонения.

Таблица 1

Изъяснительное предложение	Наклонение		Отношение индикатив/конъюнктив
	индикатив	конъюнктив	
с союзом daß	30	163	1/5
без союза daß	21	748	1/36

[5] Большой интерес может представлять и качество конечных лингвостатистических совокупностей, в частности при анализе стиля отдельного автора, при сопоставлении стилей двух или нескольких авторов, при решении задач по атрибуции спорных текстов (ср.: Головин, 1971, с. 139—143; 155—157; Войнов, 1972; Слепак, 1978, с. 12—14; и др.).

[6] См.: Ulvestad, 1962, с. 71. Форма таблицы в целях наглядности несколько изменена.

Таблица 2

Размер предложения (в словах)	Общее число предложений	Число предложений с зарамочными членами	Доля предложений с зарамочными членами в ‰
4	20	0	0
8	101	4	4.0
16	155	40	25.8
24	166	60	36.6
32	110	43	39.0

Другой пример статистической группировки — таблица 2, в которой раскрывается связь между размером предложения и выносом его членов за глагольно-сказуемную рамку в немецком языке.[7]

Совершенно ясно: чем длиннее предложение, тем чаще его члены находятся вне глагольно-сказуемной рамки. Другими словами: постановка членов предложения в зарамочную позицию существенным образом зависит от размера предложения.

Для всестороннего изучения статистической совокупности обычно требуется значительное число группировок. Так, например, О. Б. Сиротинина при исследовании факторов, определяющих различные варианты словорасположения в русском языке, группирует собранный материал то в зависимости от функционального стиля (разговорный, научный, художественный), то в зависимости от структуры сказуемого (простое, составное именное, составное глагольное), то в зависимости от коммуникативного членения предложения, то в зависимости от лексического наполнения сказуемого и т. д. (Сиротинина, 1965).

Оценка достоверности результатов качественно-количественного исследования опирается на целый ряд критериев. Основной критерий — соответствие между качественным анализом и количественными данными, внутренняя логика цифр. Так, прежде всего именно этот критерий убеждает в правомерности как вывода Б. Ульвестада о зависимости употребления наклонения в изъяснительном предложении от наличия/отсутствия союза daß, так и вывода Р. Рата о зависимости использования зарамочной позиции от размера предложения. В самом деле, отсутствие союза daß в изъяснительном предложении лишает его формальной специфики: подчиненное предложение начинает совпадать по форме с сочиненным. В этих условиях вполне естественно усиление другого формального признака подчинения, в частности конъюнктива. Аналогичным образом обстоит дело и с выводом Р. Рата. Достаточно прочесть несколько предложений с различным объемом рамки, чтобы убедиться в том, что чем больше рамка, тем труднее понимание предложения. В связи с этим не приходится удивляться тому факту, что с увеличением размера предложения растет число случаев постановки слов в зарамочную позицию, тем самым ликвидируется угроза чрезмерного расширения рамки.

[7] См.: Rath, 1965, с. 220. Форма таблицы несколько изменена.

В качестве вспомогательного критерия оценки достоверности можно использовать соотношение между сопоставляемыми цифрами. Если изучаемый фактор имеет количественное выражение, то вывод будет тем надежнее, чем отчетливее пропорциональность — прямая или обратная — между мерой фактора и мерой следствия (см., например, табл. 2, где с увеличением размера предложения последовательно растет процент предложений с зарамочными членами). Если изучаемый фактор имеет качественное выражение (например, наличие/отсутствие союза daß), то вывод будет тем надежнее, чем резче разрыв между цифрами, показывающими действие различных факторов (если бы в табл. 1 отношение индикатив/конъюнктив в изъяснительных предложениях без союза daß было не 1/36, а, допустим, 1/7, то вывод Б. Ульвестада был бы сомнителен). В тех случаях, когда соотношение между сопоставляемыми цифрами выглядит не вполне убедительно, целесообразно провести контрольный эксперимент. Нельзя, далее, упускать из виду, что надежность выводов существенным образом зависит от качества и количества отобранного для исследования материала. Материал должен быть разнообразен, в противном случае исследователь рискует упустить какие-то важные стороны изучаемого явления.

Если исследуется конечная совокупность или бесконечная однородная совокупность, то появляется возможность оценить степень достоверности вероятностно-статистическими методами. Однако применение этих методов здесь не всегда целесообразно, ибо оно требует большой затраты дополнительного труда (прежде всего в силу необходимости увеличения объема выборочного материала в несколько, иногда во много раз). Но если другие способы оценки достоверности недостаточно убедительны, использование вероятностно-статистических расчетов будет вполне оправдано.

В качественно-количественном исследовании очень труден вопрос о необходимом объеме выборки. Наиболее важный способ исходит из степени соответствия между качественным анализом и количественными данными. Крупный русский статистик А. А. Кауфман описывает его следующим образом: «Если — такова общая характеристика этого способа — полученные числа дают правильно располагающиеся ряды, притом такие, колебания которых поддаются рациональному объяснению, если, говоря короче, в цифрах обнаруживается известный внутренний смысл — значит, число наблюдений было достаточно велико, и выводы из них могут притязать на достаточную достоверность. Если, напротив, полученные числа обнаруживают колебания, лишенные правильности, если выводы резко противоречат тому,

что, по всей совокупности обстоятельств, должно было бы иметь место — возникает основание предполагать, что число наблюдений недостаточно. Это предположение подтвердится, если при увеличении числа наблюдений получатся ряды иного характера, более соответствующие тому, что мы представляем себе нормальным в данных условиях. Если, наоборот, характер колебаний цифр и при значительно большем числе наблюдений останется тот же — значит, дело не в недостаточности числа наблюдений, значит, наши выводы, основанные на цифрах, были правильны, а те наши соображения, которые оказались в противоречии с цифрами, требуют изменения или поправок» (Кауфман, 1928, с. 123). Данный способ, однако, применим только после того, как материал собран, обработан и проанализирован. На начальной стадии исследования следует учитывать структуру единиц наблюдения, их семантический потенциал и сферу функционирования: в общем, чем сложнее структура единиц наблюдения, чем богаче их семантический потенциал, чем шире сфера их функционирования, тем больше должна быть масса наблюдений.

При изучении конечной или бесконечной однородной совокупности необходимый объем выборки можно рассчитать с помощью математической статистики.

Качественно-количественное исследование обычно начинается с предварительной стадии — с формулирования гипотез на базе отрывочных наблюдений, речевого опыта, теоретических соображений. Эти гипотезы затем проверяются. Но результаты исследования обычно не исчерпываются ответом на первоначально выдвинутые гипотезы. Уже в процессе самого исследования раскрываются новые стороны изучаемых явлений и возникают новые задачи.

Многие лингвисты останавливаются на предварительной стадии исследования, полагая, очевидно, что теоретические соображения в сочетании с интуицией и отрывочными наблюдениями гарантируют достоверность выводов. Но в действительности это не так. Вот один пример. Ж. Фурке в своей «Немецкой грамматике» утверждает, что конъюнктив в изъяснительных придаточных предложениях невозможен, если глагол-сказуемое главного предложения стоит в 1-м лице настоящего времени, ибо — аргументирует он — такое придаточное предложение «передает мысль того лица, которое говорит (1-е лицо), в тот момент, когда оно говорит (настоящее время): ich glaube, du bist krank (du seist krank было бы здесь невозможно)» (Fourquet, 1952, с. 188). Статистическое исследование Б. Ульвестада убедительно показало несостоятельность данного утверждения. Б. Ульвестад собрал 184 аналогичных примера, в 34 из них (т. е. в 18% всех случаев) оказался коиъюнктив (Ulvestad, 1962, с. 64—66).

Качественно-количественные методы направлены в первую очередь на изучение качественной стороны статистических совокупностей, но они дают также возможность судить и об их количественной стороне. Особенно важную, незаменимую роль играют они при изучении статистических закономерностей в бесконечных неоднородных совокупностях, ибо здесь, как мы видели, математическая статистика бессильна. В основу количественных оценок (или прогнозов) кладется знание основных факторов, управляющих функционированием той или иной языковой единицы. Так, если установлено, что подчинение предложений выражает логические связи гораздо более дифференцированно и четко, чем сочинение, и если, далее, известно, что научное рассуждение в высшей степени нуждается в дифференцированных и однозначных средствах языкового выражения, то правомерно будет сделать вывод, что в научном стиле подчинение будет представлено значительно шире, чем сочинение. Статистические обследования конкретного материала подтверждают такой прогноз и уточняют его. Но лингвисту приходится здесь довольствоваться приблизительными оценками: «часто», «редко», «как правило», «в большинстве случаев», «примерно в 70—90%» и т. д. Объясняется это самой природой неоднородной совокупности: комбинации основных факторов на разных ее участках не совпадают и, что самое главное, не могут быть в точности предусмотрены. Поэтому, как уже отмечалось, лингвисты предпочитают говорить не о статистических закономерностях, а о тенденциях.

Серьезные различия между вероятностно-статистическими и качественно-количественными методами не исключают возможности их сотрудничества. Как мы уже отмечали, вероятностно-статистическое изучение качества лингвостатистических совокупностей оказывает помощь в разграничении разных типов речи и раскрытии действующих там факторов; при определенных условиях математическая статистика может быть использована для определения необходимого объема выборки и оценки достоверности полученных результатов. Но, с другой стороны, и математическая статистика нуждается в поддержке качественно-количественных методов, в частности, при решении вопроса об однородности лингвостатистической совокупности: вероятностно-статистические критерии здесь необходимы, но недостаточны.

Н. А. Козинцева, А. Г. Козинцев

О ВЫЯВЛЕНИИ ОБЩИХ ЗАКОНОМЕРНОСТЕЙ В РАЗНОРОДНЫХ ТЕКСТАХ

При изучении функционирования языковых единиц в текстах возникает необходимость количественной оценки наблюдаемых фактов. Если исследуемое языковое явление (например, та или иная грамматическая форма или синтаксическая конструкция) редко встречается в текстах, то лингвисты, стремясь увеличить выборку, обычно прибегают к объединению разных текстов. Цель настоящей работы — показать, как это объединение влияет на результаты исследования и как получаемая совокупность текстовых выборок может быть использована для установления общих количественных закономерностей.

Рассмотрим в качестве примера функционирование видовременных форм армянского глагола в неопределенно-личных предложениях. Нас интересует, какие частные значения видовременных форм реализуются в неопределенно-личных предложениях чаще, а какие реже.

Материалом для данной статьи служат произведения современной художественной прозы на армянском языке.[1]

По-видимому, каждый из изучаемых нами текстов в какой-то мере представляет литературу в целом. Основываясь только на этой общей предпосылке и не исследуя вопроса о степени сходства и различия между текстами, лингвисты часто подходят к разным текстам как к случайным выборкам из одной генеральной совокупности (этой совокупностью в нашем случае можно считать армянскую художественную литературу конца XIX—XX вв.). Если такой подход правилен, то следует объединить все выборки и по суммарным данным вычислить средние показатели встречаемости — т. е. взвешенные частоты — различных значений (табл. 1, столбец «взвешенные средние проценты»).

[1] Список использованных источников см. в конце статьи.

Таблица 1

Частоты частных видовых значений в неопределенно-личных предложениях в разных армянских текстах

Частные видовые значения	Раффи	Зорьян	Петросян	Ханзадян	Саинян	Рассказы I	Рассказы II	φ^2	Взвешенные средние %	Невзвешенные средние %
Актуальное и историческое	1 (0.8)	14 (10.8)	3 (1.1)	8 (2.1)	5 (5.0)	2 (4.0)	20 (10.7)	0.59	5.1	5.4
Неактуальное	18 (14.3)	54 (41.5)	14 (19.2)	113 (30.1)	23 (23.0)	14 (28.0)	49 (26.2)	0.07	27.4	26.0
Конкретно-процессное	9 (7.1)	5 (3.9)	1 (1.4)	16 (4.3)	10 (10.0)	4 (8.0)	5 (2.7)	0.24	4.8	5.3
Постоянно-непрерывное и неограниченно-кратное	41 (32.5)	6 (4.6)	3 (4.1)	23 (6.1)	14 (14.0)	11 (22.0)	23 (12.3)	0.59	11.6	13.7
Перфектное	0 (0)	8 (6.1)	12 (16.4)	50 (13.3)	6 (6.0)	2 (4.0)	14 (7.5)	0.31	8.6	7.6
Неперфектное	3 (2.4)	16 (12.3)	8 (11.0)	61 (16.3)	5 (5.0)	0 (0)	22 (11.8)	0.23	11.0	8.4
Отдельного факта	12 (9.5)	4 (3.1)	6 (8.2)	32 (8.5)	13 (13.0)	5 (10.0)	8 (4.3)	0.14	7.7	8.1
Повествовательное	42 (33.3)	23 (17.7)	26 (35.6)	72 (19.2)	24 (24.0)	12 (24.0)	46 (24.6)	0.06	23.5	25.5
Сумма	126 (99.9)	130 (100.0)	73 (100.0)	375 (99.9)	100 (100.0)	50 (100.0)	187 (100.1)	—	99.9	100.0
φ^2	0.69	0.23	0.23	0.10	0.15	0.26	0.09	—	—	—

Примечание. Здесь принята классификация частных видовых значений, разработанная в исследованиях по русскому языку (см., например: Бондарко, 1971) и использованная в работе: Козинцева, 1977. Из-за недостаточного количества материала мы объединяем актуальное и исторические, значения привычного, а также постоянно-непрерывное и неограниченно-кратное значения перфекта. Приводятся абсолютные численности и (в скобках) проценты.

Действительно ли, однако, можно считать наши тексты случайными выборками из одной совокупности? Для решения этого вопроса в лингвистике уже давно применяется критерий χ^2. Результат (216.74 при 42 степенях свободы) свидетельствует о том, что в пользу гипотезы о принадлежности наших выборок к одной статистической совокупности не имеется даже одного шанса из 1000. В последней строке и в третьем от конца столбце таблицы 1 приведены показатели относительного различия ($\varphi^2=\chi^2/n$), позволяющие судить о том, в какой строке колебания частот наиболее сильны и какие столбцы в наибольшей степени отличаются от других. Из всех частных значений самая непостоянная встречаемость в разных текстах у актуального и исторического значений презенса, а также у постоянно-непрерывного и неограниченно-кратного значений имперфекта (φ^2 в обоих случаях равно 0.59). Из всех текстов наибольшим своеобразием отличается текст Раффи ($\varphi^2=0.69$).

Установив, что исследуемые тексты не принадлежат к одной совокупности, мы сразу же делаем практический вывод: суммирование численностей и выведение взвешенных средних частот недопустимо. Ведь объемы разных выборок неодинаковы (от 50 примеров из Рассказов I до 375 из произведения Саинян) и, главное, совершенно непропорциональны встречаемости соответствующих текстов в общем корпусе текстов данного периода. Достичь выборочным путем такой пропорциональности — задача чрезвычайно трудная, и мы ее перед собой не ставили. Следовательно, взвешенные частоты определяются случайным и несущественным фактором — объемами выборок. Так, если бы число примеров из Раффи оказалось бо́льшим, то взвешенная частота постоянно-непрерывного и неограниченно-кратного значений возросла бы, а частота актуального и исторического уменьшилась; увеличение доли текстов Зорьяна привело бы к обратному сдвигу; и т. д.

Таблица 2

Ранги частот видовых значений в неопределенно-личных предложениях в разных армянских текстах

Частные видовые значения	Раффи	Зорьян	Петросян	Ханзадян	Саинян	Рассказы I	Рассказы II	Сумма рангов
Актуальное и историческое	7	4	6.5	8	7.5	6.5	5	44.5
Неактуальное	3	1	2	1	2	1	1	11
Конкретно-процессное	5	7	8	7	5	5	8	45
Постоянно-непрерывное и неограниченно-кратное	2	6	6.5	6	3	3	3	29.5
Перфектное	8	5	3	4	6	6.5	6	38.5
Неперфектное	6	3	4	3	7.5	8	4	35.5
Отдельного факта	4	8	5	5	4	4	7	37
Повествовательное	1	2	1	2	1	2	2	11

Теперь поставим другой вопрос: есть ли что-нибудь общее между всеми текстами? В случае отрицательного ответа на этот вопрос подсчет каких бы то ни было средних частот не будет иметь ни малейшего смысла, так как для подавляющего большинства текстов эти частоты окажутся совершенно нехарактерными. Однако рассмотрение процентов в таблице 1 показывает, что общие черты есть. Во всех семи выборках повествовательное значение аориста занимает по своей встречаемости либо первое, либо второе место. В шести выборках то же относится к неактуальному значению презенса (лишь у Раффи оно оказывается на третьем месте). Это именно те значения, для которых показатель φ^2 минимален — соответственно 0.06 и 0.07. Ранжируем так же (обратно пропорционально частоте) и все остальные значения во всех текстах, а затем суммируем ранги для каждого значения (см. табл. 2).

Для проверки важно, что если имеется n столбцов и k строк, то общая сумма рангов должна равняться $nk\,(k+1)/2$, в нашем случае $7 \cdot 8 \cdot 9/2=252$.

Если бы никакой общей закономерности в разных текстах не было, сумма рангов в каждой строке была бы близка к средней величине (252/8=31.5). В действительности наблюдаются колебания от 11 до 45. Для проверки значимости этих колебаний (т. е. существенности различий между строками) используется критерий Фридмэна [Friedman, 1937; Закс, 1976], не нашедший еще широкого применения в лингвистике:

$$\chi_r^2=\frac{12}{nk\,(k+1)}\sum^{k}R^2-3n\,(k+1),$$

где R — сумма рангов в строке. Величина χ_r^2 распределена приблизительно как χ^2 с $(k-1)$ степенями свободы. Вычисляем:

$$\chi_r^2=\frac{12}{7\cdot 8\cdot 9}\,(44.5^2+11^2+\ldots+11^2)-3\cdot 7\cdot 9=30.74,$$

что при 7 степенях свободы свидетельствует о высшей степени существенности различий ($P<0.001$). Тот же результат можно получить по формуле Пэйджа:

$$\chi_r^2=\frac{6\Sigma\,(R-E)^2}{\Sigma R},$$

где E — средняя сумма рангов (в нашем случае — 31.5). Соответственно:

$$\frac{6\,[(44.5-31.5)^2+(11-31.5)^2+\ldots+(11-31.5)^2]}{252}=30.74.$$

Следовательно, несмотря на то что наши тексты нельзя считать случайными выборками из одной совокупности, в них прослеживается отчетливая общая закономерность: всюду частота неактуального и повествовательного значений высока, а частота конкретно-процессного, а также актуального и исторического значений низка. Прочие значения по своей встречаемости занимают промежуточное место.

Поскольку общие закономерности найдены, допустимо и вычисление общих показателей. В этом пункте мы расходимся с В. А. Никоновым [Никонов, 1978]. Справедливо критикуя авторов, игнорирующих различия между текстами, он приходит к следующему выводу: «Поэтому „среднестатистические" по языку пока нереальны. Вместо них необходимы подсчеты, дифференцированные по видам речи» [Никонов, 1978, с. 106]. Это верно лишь в том случае, если никаких общих закономерностей в текстах нет. Если же статистически доказано, что какое-то явление во всех текстах встречается редко (пусть неодинаково редко), а какое-то явление всюду гораздо чаще (пусть неодинаково часто), то мы не видим причин не указать — хотя бы в ориентировочных целях — частоту, вокруг которой колеблется встречаемость каждого из явлений.

Другой вопрос — как вычислить эту среднюю частоту. Как мы убедились, взвешенные средние частоты — показатели в данном случае неудовлетворительные. Зато альтернативный способ — подсчет невзвешенных средних частот — по-видимому, приемлем, особенно если в дополнение к средним частотам приводить пределы колебаний встречаемости явления в разных текстах. Действительно, раз мы не можем сделать объемы выборок пропорциональными встречаемости данных текстов в литературе, то приходится вообще не учитывать размеры выборок (если, конечно, они не слишком малы) и приписывать всем текстам одинаковый вес. Конечно, надежность полученных таким способом частот не следует переоценивать: ведь гипотеза об одинаковой представленности текстов в общем корпусе априори ошибочна.

Более точные оценки средних частот можно было бы получить, применив принцип случайного отбора материала [см.: Кокрен, 1976]. Однако, во-первых, такой принцип весьма трудно осуществим в лингвистике, а во-вторых — и это главное — сама задача оценки параметров суммарной совокупности в значительной мере теряет актуальность ввиду различий между текстами. Обнаружить общие тенденции в таком случае важнее, чем вычислить средние показатели.[2]

[2] «. . . Характеристики соотношения между употребительностью падежей. . . оказываются более точными, чем какая-нибудь характеристика, устанавливающая жесткие верхние и нижние границы колебаний в употребительности этих падежей. ., т. е. рассматривающая эти колебания как несущностные вариации», — правильно отмечает В. Г. Адмони [Адмони, 1970, с. 98]. Сходные мысли высказывает и В. А. Никонов [Никонов, 1978, с. 106]. Задача поэтому состоит в том, чтобы, зафиксировав в пределах речи «сущностные

Невзвешенные (средние арифметические) частоты приведены в последнем столбце таблицы 1. Они не очень отличаются от взвешенных — максимум на 2—3%. Не нужно, однако, думать, что такое соответствие будет наблюдаться всегда. Дело в том, что, во-первых, максимальные размеры выборок приходятся у нас на тексты с минимальным значением φ^2 — Рассказы II и произведения Ханзадяна, а во-вторых, благодаря подбору довольно разно-

Таблица 3

Зависимость оценки суммарной частоты от результатов тестов на различие и на сходство между выборками

Тест на различие между выборками (χ^2)	Тест на сходство между выборками (критерий Фридмэна при $n > 2$ и коэффициент ранговой корреляции при $n = 2$)	Интерпретация	Оценка суммарной частоты
+	+	*Выборки не принадлежат* к одной совокупности, но обнаруживают общую закономерность	*Средняя не*взвешенная
+	−	Выборки не принадлежат к одной совокупности и не обнаруживают общей закономерности	Не существует
−	+	Выборки могут принадлежать к одной совокупности и обнаруживают общую закономерность	Средняя взвешенная
−	−	Выборки могут принадлежать к одной совокупности, но не обнаруживают общей закономерности. Такой случай скорее всего объясняется тем, что выборки слишком малы. Другое объяснение (имеющее, правда, лишь теоретический интерес): выборки принадлежат к одной совокупности, в которой частота всех изучаемых единиц одинакова.	Средняя взвешенная, имеющая в этом случае лишь очень ограниченную ценность

Примечание. Плюс означает, что значение критерия превышает критический уровень, минус ставится при недостоверном результате.

вариации» (различия между текстами), найти все же такие закономерности, которые свойственны значительному большинству текстов и соответственно речи в целом. Ранговые критерии представляются перспективным средством решения этой задачи.

образных текстов, «крайности» уравновесились. Последнее обстоятельство особенно важно. При комплектовании выборок следует, с одной стороны, добиваться максимальной внутритекстовой однородности (всячески избегая объединения текстов, имеющих разные закономерности), но, с другой стороны, стремиться к наибольшей межтекстовой разнородности, повышая тем самым репрезентативность общей совокупности. Иными словами, необходимо подбирать тексты внутренне единые, но контрастирующие между собой. Это, по-видимому, не приведет к значительному искажению частот, свойственных суммарной совокупности, но усилит пестроту материала. Чем контрастнее окажутся тексты, тем выше вероятность того, что обнаруженная в них общая закономерность свойственна речи в целом.

Достичь внутреннего единства каждой из выборок далеко не просто. Использование текстов одного автора отнюдь не всегда дает желаемый результат. Мы разделили выборку из произведений Ханзадяна на две части — относящиеся соответственно к романам «Земля» (149 случаев) и «Каджаран» (226 случаев). Оказалось, что эти две подвыборки весьма существенно различаются по распределению частных значений в неопределенно-личной конструкции ($\chi^2=19.34$, при 7 степенях свободы $P<0.01$), хотя и ранговый коэффициент корреляции между обеими подвыборками (показатель, аналогичный критерию Фридмэна и применяемый при $n=2$) существен ($r_s=+0.89$, $P<0.01$). Возможно, и в пределах одного произведения не будет наблюдаться однородности. Не исключено, что внутритекстовые различия в функционировании рассматриваемых языковых единиц (например, между авторской речью и диалогом) окажутся даже сильнее, чем различия между текстами разных авторов. Однако критерии для выделения функционально однородных (в рассматриваемом аспекте) частей текста находятся пока в стадии разработки.

В заключение сведем все четыре случая, возможных при использовании тестов на различие и на сходство между выборками, в таблицу 3.

СПИСОК СОКРАЩЕНИЙ ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ (на армянском языке)

Зорьян — З о р ь я н С. 1) История одной жизни. Ереван, 1955; 2) Первые дни. Ереван, 1967; 3) Дед и внук. Яблоневый сад. Мать. Царь Пан (отрывок). — В кн.: Избранные страницы армянской литературы. Ереван, 1946; 4) Оганес. Забор. Неизвестные нити. Невестка Закара. — В кн.: Армянская новелла. Ереван, 1971.

Петросян — П е т р о с я н Р. Цахкаланч. Ереван, 1969.

Рассказы I — Прозаические произведения конца XIX—начала XX в. следующих писателей: Г. Агаяна, Раффи, Мурацана, Ширванзаде, Лео, Л. Манвеляна, Атрпета, Нар-Доса, О. Туманяна, Г. Башинджагяна, В. Папазяна и А. Исаакяна, опубликованные в кн.: Избранные страницы армянской литературы. Ереван, 1946; Армянская новелла. Ереван, 1971.

Рассказы II — Прозаические произведения писателей советского периода: Д. Демирчяна, М. Манвеляна, Араази, Аракса, В. Ананяна, Г. Беса, Г. Севунца, В. Тотовенца, А. Бакунца, Г. Кочара, опубликованные в тех же сборниках, а также следующие повести и рассказы: С е в у н ц Г. Иско из Оцанвайра. — Советакан граканутюн, 1971, № 3; Г у р у н ц Л. Гусар Мухан. — Советакан граканутюн. 1970, № 3; О в с е п я н Р. Голоса. — Советакан граканутюн, 1970, № 6; А м и р х а н я н А. Звезды. . . — Там же; Б а л а я н В. Уголовное дело №. . . — Там же; А р а м я н В. Рассказы. . . — Там же; Ш а к а р я н Ф. Ты мой друг. — Там же.

Раффи — Р а ф ф и. Безумный. — В кн.: Раффи. Безумный. Гарем. Золотой петушок. Рассказы. Ереван, 1970.

Саинян — С а и н я н А. Пути-дороги. Ереван, 1958.

Ханзадян — Х а н з а д я н С. Н. Собр. соч. в пяти томах. Т. 2. Ереван, 1967; т. 5, Ереван, 1970.

А. В. Грошева

СТАТИСТИЧЕСКАЯ ОДНОРОДНОСТЬ ТЕКСТОВ НА СИНТАКСИЧЕСКОМ УРОВНЕ

Проверка однородности текстов относительно частот лингвистических объектов — кардинальный вопрос лингвостатистики. От чисто теоретических рассуждений о том, может ли литературный язык определенной эпохи и его различные функциональные стили обладать статистически однородными признаками или это свойственно только узко специальным текстам, ученые все больше и больше обращаются к конкретным исследованиям языкового материала. Предметом нескольких статей в сборнике «Вопросы статистической стилистики» является выяснение однородности разных по жанру текстов с точки зрения признаков лексического, лексико-грамматического и грамматического характера (см.: ВСС, 1974). На синтаксическом уровне вопрос об однородности текстов в его своеобразном преломлении — решение проблемы спорного авторства — ставился в работе Г. У. Юла (Yule, 1939). Юл математически сформулировал гипотезу, что длина предложения и ее колебание вокруг средней величины являются характеристикой авторского стиля. Результаты Юла ограничены, так как они опираются на тексты, однородные с точки зрения их построения. Анализом длины предложения, при котором учитывались факторы построения текста, занимался Г. А. Лесскис (Лесскис, 1963).

Перед нашим исследованием была поставлена задача выявить однородность текстов на синтаксическом уровне в пределах одного жанра — латинской исторической прозы. Разумеется, рассмотрение лишь одного из разнообразных функциональных стилей языка сужает исследование и является определенным недостатком. Не исключена возможность, что разные функциональные стили являются взаимоисключающими генеральными совокупностями. Однако латинская историческая проза — это лишь первый этап исследования, за которым должны последовать другие; пока мы располагаем только данными статистического

обследования нескольких произведений исторического характера, принадлежащих латинским авторам.

В работах по лингвостатистике понятие «однородность текстов» употребляется не всегда однозначно. В данной статье оно имеет весьма конкретное определение: речь пойдет об однородности текстов только в отношении двух языковых признаков, которые подвергались изучению — в отношении длины (т. е. размера, объема) цельного предложения (ЦП) и длины элементарного предложения (ЭП). Лексическая длина предложения — важный языковой фактор, один из элементов, характеризующих структуру предложения. Необходимо было выяснить, насколько связные законченные тексты однородны относительно частот двух указанных признаков синтаксического уровня.

Жанр исторической прозы имеет в латинской литературе длительную традицию. Для данного исследования были отобраны произведения следующих авторов: 1) Цезарь, «Записки о Галльской войне» (Цез.); 2) Саллюстий, «Югуртинская война» и «Заговор Катилины» (Салл.); 3) Ливий, «От основания города» (Лив.); 4) Тацит, «Анналы» (Тац.); 5) Аммиан Марцеллин, «Римская история» (Марц.).[1]

Произведения всех названных авторов, кроме Саллюстия, весьма обширны и имеют деление на книги. «Записки о Галльской войне» Цезаря, сохранившиеся в полном объеме, составляют 7 книг; из каждой книги этого произведения, начиная с 5-й главы, были произведены выборки одинаковой длины — по 200 ЭП. По типу выборок из «Записок . . .» Цезаря были осуществлены выборки из произведений других авторов. Из 35 сохранившихся книг Ливия, при соблюдении принципа случайности отбора, для обследования были взяты следующие семь: 1-я, 6-я, 21-я, 27-я, 33-я, 39-я, 45-я. Из дошедших до нас книг «Анналов» Тацита для обследования были отобраны четные тома, дополненные выборкой из 1-й книги. Из 18 сохранившихся книг исторического труда Аммиана Марцеллина выборки почерпнуты из книг 14-й, 17-й, 20-й, 22-й, 25-й, 28-й, 31-й. Небольшое произведение Саллюстия «Югуртинская война» издателями разделено на главы; I выборка охватывает гл. 5—17, II — гл. 17—28; остальные выборки разделяются равными промежутками (III — гл. 34—44, . . . VII — гл. 99—108).

Таким образом, общий объем выборки по каждому автору составляет 1400 ЭП. Предметом изучения была только авторская

[1] Выборки производились по следующим изданиям: 1) Caesaris C. Julii. Commentarii de Bello Gallico, 4-е изд. М., 1892; 2) Sallusti C. Crispi Catilina-Jugurtha. Fragmenta ampliora. Ed. 2. Lipsiae, 1954; 3) Livi T. Ab urbe condita libri. Bd. 1—10. Berlin, 1881—1911; 4) Tacitus Publius Cornelius. Annalen. Bd. 1—4. Heidelberg, 1963—1968; 5) Marcellinus Ammianus. Romische Geschichte. Lateinisch und Deutsch. Bd. 1—4. Berlin, 1968—1971.

речь; вставные речи — образцы исторического вымысла, которыми так изобилуют сочинения античных историографов — выпускались.

I. Однородность текстов на уровне ЦП

Исходными данными любого эксперимента служат наблюденные частоты основных единиц подсчета. Мы подсчитали количество слов в цельном предложении (а), в элементарном (б) и получили в свое распоряжение данные семи выборок для каждого автора. Ввиду того, что колебания в размерах цельного предложения достаточно велики (от одного слова до ста и более), для удобства произведения подсчетов цельные предложения были разбиты на классы с интервалом в семь слов. Результаты подсчетов ЦП в выборках из сочинения Ливия представлены в таблице 1.

Таблица 1

Количество ЦП в тексте Ливия

Классовый интервал	№№ классов	№№ выборок							Всего ЦП
		I	II	III	IV	V	VI	VII	
2—8	I	18	22	11	14	14	24	14	117
9—15	II	32	20	19	22	15	23	30	161
16—22	III	15	18	17	11	16	12	10	99
23—29	IV	9	9	13	6	7	9	3	56
30—36	V	4	2	6	3	5	4	8	32
37—43	VI	2	2	3	3	2	2	2	16
44—50	VII	1	1	—	2	2	2	2	10
51—57	VIII	—	1	—	0	1	0	1	3
58—64	IX	—	0	—	1	1	0	1	3
65—71	X	—	0	—	3	0	0	—	3
72—78	XI	—	1	—	—	0	1	—	2
79—85	XII	—	—	—	—	0	—	—	0
86—92	XIII	—	—	—	—	1	—	—	1
Всего ЦП		81	76	69	65	64	77	71	503

Располагая приведенными данными (табл. 1), мы имеем возможность сравнивать между собой частотные ряды попарно взятых выборок. При этом необходимо принять определенную начальную гипотезу, так называемую нулевую гипотезу. В качестве таковой выступает предположение о совпадении вероятностей. Затем это предположение проверяется с помощью статистических критериев.

Для сравнения двух наблюдаемых вероятностей в качестве статистического критерия нами был избран χ^2, поскольку он

позволяет, во-первых, сравнивать частотные ряды с неизвестным законом распределения и, во-вторых, устанавливать достаточную (или недостаточную) близость конкретных сравниваемых значений.

Так как в начале эксперимента невозможно заранее предсказать ожидаемые результаты, мы сочли необходимым вычислить критерий χ^2 для проверки однородности всех возможных парных сочетаний из семи имеющихся в нашем распоряжении выборок. Данные I выборки последовательно сопоставлялись с данными II, III, . . . VII выборок, затем данные II выборки — с данными III, IV, . . . VII выборок и т. д.; всего было произведено 21 сопоставление.

Предположим, что мы сравниваем частотный ряд из 7 чисел ($\mu_i=7$) в I выборке с аналогичным рядом из 11 чисел ($\nu_i=11$), полученным в VI выборке. Общее количество ЦП в рассматриваемых выборках не одинаково: $m=81$ и $n=77$.

Первоначально для каждого класса ЦП вычисляется выражение $\frac{1}{\mu_i+\nu_i}\left(\frac{\mu_i}{m}-\frac{\nu_i}{n}\right)^2$, составляющее индивидуальное «классовое» значение χ^2. Затем все индивидуальные значения суммируются, давая χ^2 для всего ряда классов ЦП:

$$\chi^2_{(1)}=0.000\,196$$
$$\chi^2_{(2)}=0.000\,169$$
$$\chi^2_{(3)}=0.000\,032$$
$$\vdots$$
$$\chi^2_{(7)}=0.000\,061$$
$$\vdots$$
$$\chi^2_{(11)}=0.000\,169$$
$$\overline{\Sigma\chi^2=0.000\,630.}$$

Количество степеней свободы в данном случае $k-1$ $(11-1)$ равно 10. Однако оценивать полученное значение χ^2 еще рано: необходимо привести это значение в соответствие с различными объемами выборок. Для этого суммарное значение χ^2 умножается на произведение mn, давая истинное значение наблюдаемого χ^2 (в данном случае $\chi^2=3.93$), которое затем проверяется в таблице критических значений в строке $f=10$. Наименьшее значение величины χ^2_q, превосходящее 3.93 при $k=10$, равно 3.94. Этому значению χ^2_q соответствует $q_i=0.95$ (q — процентные точки распределения χ^2, q_i — значение q, найденное в эксперименте), что заведомо гораздо выше любого стандартного уровня значимости; кстати, для нашего эксперимента берется уровень значимости, равный 0.1 (10%).

Следовательно, значение критерия χ^2, найденное в эксперименте, находится в области допустимых значений, поэтому нулевая

гипотеза о совпадении частотных распределений в двух выборках и о совпадении конкретных частот ЦП не отвергается.

Сводные результаты по применению критерия χ^2 для проверки однородности частот ЦП в двух последовательно рассматриваемых выборках из текста Ливия даны в таблице 2.

Таблица 2

Однородность частот ЦП в тексте Ливия

Выборки	$k-1$	χ^2	Оценка (в %%)
I—II	10	5.96	80
I—III	6	6.57	40
I—IV	9	6.60	70
I—V	12	8.45	75
I—VI	10	3.93	95
I—VII	8	7.53	50
II—III	10	8.82	55
II—IV	10	10.04	40
II—V	12	6.59	90
II—VI	10	3.05	98
II—VII	10	15.34	10
III—IV	9	11.37	20
III—V	12	7.75	80
III—VI	10	10.97	40
III—VII	8	15.22	5
IV—V	12	7.98	80
IV—VI	10	7.68	65
IV—VII	9	7.34	60
V—VI	12	7.95	80
V—VII	12	9.13	70
VI—VII	10	4.57	90

Для того чтобы возможно более точно оценить полученные результаты, мы разделили область допустимых значений q на так называемые зоны однородности:

I зона — $q < 99\%$, но $> 75\%$
II зона — $q < 75\%$, но $> 50\%$
III зона — $q < 50\%$, но $> 25\%$
IV зона — $q < 25\%$, но > 0.

Первую зону мы считаем зоной высокого соответствия сравниваемых величин, вторую — зоной хорошего соответствия; попадание выборок в обе эти зоны свидетельствует о большом сходстве их частот. Третья зона считается зоной среднего соответствия, четвертая — низкого соответствия.

Подсчеты показали, что в тексте Ливия из 21 пары выборок в зоне высокого соответствия оказалось 9 пар, во второй зоне — 6 пар, в третьей и четвертой — по 3 пары. Таким образом, 70% рассмотренных пар выборок из текста Ливия обнаружили большую степень однородности на синтаксическом уровне, а именно в объеме ЦП.

Важно отметить, что однородными в тексте Ливия оказались выборки, не расположенные в непосредственной близости друг от друга, например, I—VI, II—VI.

Как видим, в пределах одного произведения получены противоположные результаты: опровержение нулевой гипотезы для выборок III—VII (q_i=0.05) и очень высокое значение q_i (0.95) для выборок I—VI и даже q_i, равное 0.98, для выборок II—VI.

Эти результаты показывают, что нельзя сделать однозначный вывод об однородности или неоднородности двух произвольно взятых выборок из одного произведения относительно частот ЦП. Можно лишь отметить, что абсолютно негативный результат получился в одном случае из 21. Иными словами, если взять из одного текста две произвольные выборки, то скорее можно ожидать, что нулевая гипотеза при проверке критерия χ^2 не будет отвергнута.

Однако мы не могли удовлетвориться результатами и выводами, полученными на основании наблюдений над текстом одного автора (Ливия) и повторили эксперимент над идентичными по объему выборками из произведений других представителей латинской исторической прозы — Цезаря, Саллюстия, Тацита и Марцеллина, поскольку каждый из этих авторов обладал своей индивидуальной стилевой манерой и их творчество относится к разным хронологическим периодам.

В процессе изучения однородности текста у предвестника римской историографии Цезаря, как и при изучении текста Ливия, были получены результаты противоположного характера, как опровергающие нулевую гипотезу о совпадении частотных распределений длин ЦП (выборки из книг 2-й и 4-й, 4-й и 7-й, 3-й и 7-й, для которых $q_i \leqslant 10\%$), так и свидетельствующие о довольно значительной степени однородности двух выборок (I и III при q_i=85%; V и VI при q_i=75%).

Остановимся более подробно на вопросе стабильности частот ЦП в «Записках о Галльской войне» Цезаря. Важно отметить большое колебание в самом количестве ЦП, взятых из отрезков, равных по числу ЭП: от 54 в 4-й книге до 85 — в 6-й. Это различие является отражением разной степени насыщенности цельных предложений элементарными: коэффициент комплексности (КК) в 4-й книге равен 3.7, в 6-й книге — 2.35. Однако эти полярные по содержанию в них цельных предложений выборки отнюдь не оказались неоднородными, дав значение χ^2, равное 13.23, которому при k=11 соответствует q_i=0.25, т. е. выше допустимого уровня значимости.

Среди специалистов существует мнение, что из семи книг «Записок . . .» Цезаря книги со 2-й по 6-ю (т. е. центральная часть произведения) относятся к числу наиболее отшлифованных, а 1-я и 7-я якобы не подверглись литературной отделке и представляют собой черновые наброски (Dernoschek, 1903). В 1-й и 7-й книгах Цезарь рассказывает о «наиболее крупных и угрожающих событиях» (История римской литературы, 1959), а именно

о войне с германским вождем Ариовистом, вторгшимся в пределы Галлии (кн. 1-я), и о последнем огромном восстании всех галльских племен под предводительством Верцингеторига (кн. 7-я). Остальные книги повествуют о непрерывных столкновениях римлян с галльскими и германскими племенами.

Каковы бы ни были особенности 1-й книги «Записок . . .», выборка I в отношении частот ЦП обнаружила неплохое соответствие со всеми остальными выборками (совокупность q_i=325%); лучшее соответствие имеет только выборка из 5-й книги (совокупность q_i=350%). Значит, с точки зрения частоты длин ЦП 1-я книга не противопоставляется 2-й—6-й книгам. Зато текст из 7-й книги действительно оказался наименее однородным в сопоставлении с выборками из предыдущих книг (совокупность q_i=135%), дав в пяти случаях из шести такие значения q_i, которые, хотя и лежат в области допустимых, но весьма малы; а в двух случаях, как было указано выше, анализ выявил неоднородность выборок III—VII, IV—VII. Степень однородности 7-й книги с 1-й также невелика (q_i=0.4), что ставит под сомнение ее одинаковое с 7-й книгой положение в отношении стилевой необработанности.

Хотя исследователями стиля Цезаря нигде не отмечалось особое положение 4-й книги среди других книг «Записок о Галльской войне», однако при изучении длины ЦП в отрывке из этой книги выяснилось, что она написана предельно сложными даже для Цезаря предложениями. По-видимому, этот факт сыграл решающую роль в том, что однородность 4-й книги со всеми остальными книгами оказалась довольно низкой (совокупность q_i=175%) и по степени однородности между 4-й и, например, 6-й книгами (совокупность q_i=250%) намечается большой разрыв. Разрывы в совокупности q_i между остальными выборками не столь велики (II — 270%, III — 295%, I — 325%, V — 350%).

Конечно, на основании только одного эксперимента с одним языковым признаком (длина ЦП) нельзя делать далеко идущих выводов, однако возможно, что выявление плохого соответствия частот ЦП в 7-й книге с частотами ЦП в других книгах свидетельствует в пользу версии, будто эта книга не подвергалась столь тщательной отделке, как 2-я—6-я книги, и носит характер чернового наброска. В то же время наш эксперимент отвергает подобную версию для 1-й книги.

Чтобы не перегружать статью фактическими данными, для остальных авторов мы приведем лишь сведения о распределении 21 пары выборок по зонам однородности в сравнении с уже известными показателями однородности частот в тексте (см. табл. 3).

Как видно из таблицы 3, картина однородности выборок очень пестра и у каждого автора имеет свои особенности. Примерно одинаковое количество однородных пар выборок из текстов Саллюстия, Ливия и Тацита располагается в двух верхних зонах. Но если у Ливия преобладает зона высокого соответствия —

Таблица 3

Распределение сопоставляемых выборок по зонам однородности на уровне ЦП

Зона	Автор					Пары выборок
	Цез.	Салл.	Лив.	Тац.	Марц.	
I	2	7	9	6	7	31
II	6	7	6	9	3	31
III	8	6	3	2	5	24
IV	5	1	3	4	6	19
Пары выборок	21	21	21	21	21	105

I, то у Тацита — II, т. е. процентное выражение однородности выборок у него ниже, чем у Ливия. У Саллюстия выборки распределены поровну между I и II зонами. Если сравнить данные двух верхних зон у Цезаря и Ливия, то совершенно очевидно, что это сравнение не в пользу Цезаря: в зонах высокого и хорошего соответствия у него оказалось только 38% всех пар выборок по сравнению с 70% у Ливия. Что касается III и IV зон, здесь у наших авторов также не наблюдается единства: у Саллюстия преобладают выборки со средним соответствием и лишь одна находится в зоне низкого соответствия (IV—VI, $q_i < 10\%$, т. е. выборки неоднородны). У Тацита 4 пары выборок оказались в зоне низкого соответствия, из них I—VI неоднородны. У Марцеллина также своеобразное распределение выборок: число выборок, оказавшихся в зоне высокого соответствия довольно велико (1/3 общего количества), но почти столько же выборок находится и в зоне низкого соответствия, из них 2 пары явно неоднородны, а 3 пары находятся на грани допустимых значений q_i.

Несмотря на пестроту картины, некоторые общие для всех авторов выводы все-таки можно сделать. Из 105 сопоставлений около 1/3 выборок обнаружили высокую степень однородности и столько же — хорошую степень. Низкий уровень однородности засвидетельствован в наименьшем количестве выборок, всего в 17%, из них абсолютно неоднородными оказались 8 пар выборок, т. е. менее 10%. У каждого из исследуемых авторов хотя бы одна пара выборок оказалась неоднородной.

Метод распределения выборок по зонам однородности хотя и позволяет делать интересные наблюдения, однако не дает ответа на вопрос, тексты какого из авторов отличаются наиболее высокой степенью однородности. Тем не менее, как нам кажется, ответ на этот вопрос можно получить самым простым способом, суммировав процентные выражения q_i по 21 паре сопоставляемых выборок для каждого автора (см. табл. 4).

Таблица 4

Степень однородности текстов относительно ЦП

Автор	Цез.	Салл.	Ливий.	Тац.	Марц.
q_i (в %%)	900	1200	1293	1191	1011

Отсюда с очевидностью следует, что наибольшей однородностью относительно частот длин ЦП отличаются текст Ливия, затем тексты Саллюстия и Тацита, потом Марцеллина. Как ни странно, последнее место в этом перечне занимает текст Цезаря.

Необходимо выяснить также вопрос о том, какие именно классы ЦП обнаруживают при сопоставлении выборок лучшее соответствие частот и какие — худшее. Дело в том, что выборки имеют неодинаковое количество классов, и рассматривать следует только те классы, в которых есть данные по 21 паре сопоставляемых выборок. Назовем такие классы полными. Так, у Цезаря и Марцеллина в области ЦП 9 полных классов, у Тацита — 7, у Саллюстия и Ливия — по 6. В таблицах 5 и 6 представлены классы, обнаружившие лучшее или худшее соответствие при сопоставлении выборок.

Таблица 5

Классы ЦП с лучшим соответствием частот *

Автор	Цез.	Салл.	Ливий.	Тац.	Марц.
Класс	V	III	VI	VI	VII
Частота	12	104	16	22	41
Совокупное значение χ^2	2582	1821	644	1139	1130
Модальный класс	II—III	II	II	III	III—IV

* В этой таблице и ниже приводятся только значащие цифры совокупного классового значения χ^2; в действительности оно равняется, например, у Цезаря 0.002 582, у Саллюстия 0.001 821 и т. д.

Таблица 6

Классы ЦП с худшим соответствием частот

Автор	Цез.	Салл.	Лив.	Тац.	Марц.
Класс	II	V	IV	I	II
Частота	103	36	56	39	56
Совокупное значение χ^2	8531	6631	4904	6493	10 982

Как явствует из таблиц, у всех авторов, кроме Саллюстия, наблюдается определенная тенденция: худшее соответствие в классах с меньшим размером ЦП и большей частотой, лучшее соответствие в классах с большим размером ЦП и меньшей частотой. Только у Саллюстия обратные отношения, но это безусловно объясняется тем, что у этого автора вообще нет больших различий между индивидуальными классовыми значениями χ^2:

Классы	III	IV	I	VI	II	V
Значение χ^2	1821	1977	2336	2383	2925	6631

Только у Цезаря класс с худшим соответствием частот является одним из модальных, т. е. самых высокочастотных классов. Так что, казалось бы, для этого автора будет правильным утверждение, что класс ЦП самой высокой частоты является наименее стабильным. Но это утверждение опровергается данными III класса, который является модальным, как и II, однако обнаруживает по всем выборкам гораздо лучшее соответствие частот (совокупное значение χ^2=0.003312). Следовательно, устойчивость или неустойчивость частоты классов ЦП нельзя непосредственно и однозначно связывать с большой величиной частоты какого-либо класса.

Следующим этапом нашего эксперимента было сравнение между собой нескольких текстов разных авторов. Сравнивались последовательно выборки из «Записок . . .» Цезаря с отрезками текста Саллюстия, Ливия, Тацита и Марцеллина (см. табл. 7); затем выборки из текста Саллюстия с выборками из произведений Ливия, Тацита, Марцеллина и т. д. (см. ниже табл. 8, 9, 10). Чтобы облегчить расчеты, были взяты для сравнения выборки с одинаковым количеством ЦП — единственное несущественное ограничение, которое не может повлиять на результаты сопоставления текстов двух различных авторов.

Только тексты Цезаря и Тацита оказались в большой степени однородными (см. табл. 7): одна пара выборок расположена в зоне высокого соответствия (обе из начала произведений), другая пара — в зоне хорошего соответствия, третья — в зоне среднего. Привлечение дополнительно еще двух пар выборок из текстов Цезаря и Тацита (II—IV и VII—II) мало изменило картину: значения q_i оказались хотя и довольно низкими (20% и 25%), но не отвергающими гипотезу об однородности выборок. Сравнение текстов Цезаря и Марцеллина дало результаты противоположного характера. Следовательно, выборки могут оказаться как однородными, так и неоднородными. Сопоставление же текста Цезаря с текстами его ближайших современников

Таблица 7

Сравнение текста Цезаря с текстами других авторов

Авторы	1	Цез.—Салл.			Цез.—Лив.		
Выборки	2	V—I	VII—III	III—IV	VII—VII	V—II	V—VI
χ^2	3	16.86	14.81	20.48	17.55	16.5	22.34
$k-1$	4	11	8	14	8	11	11
q_i (в %%)	5	12	7	12	2.5	12	2

Таблица 7 (*продолжение*)

Авторы	1	Цез.—Тац.			Цез.—Марц.		
Выборки	2	I—I	V—V	II—III	II—II	V—IV	I—V
χ^2	3	8.13	12.81	10.91	9.83	20.07	21.62
$k-1$	4	12	11	12	12	12	12
q_i (в %%)	5	87	30	55	63	7	4

дало такие значения χ^2, которые либо находятся на грани достоверных значений, либо ниже их.

Итак, относительно частоты ЦП различной длины тексты Цезаря и Саллюстия, Цезаря и Ливия следует считать неоднородными; тексты Цезаря и Тацита обнаруживают хорошее или низкое соответствие частот ЦП; тексты Цезаря и Марцеллина не имеют устойчивого соответствия частот.

Таблица 8

Сравнение текста Саллюстия с текстами других авторов по частоте длин ЦП

Авторы	Салл.—Лив.			Салл.—Тац.			Салл.—Марц.	
Выборки	I—II	III—VII	VI—I	I—V	III—II	II—III	II—IV	I—III
χ^2	9.07	7.28	2.56	14.56	6.38	12.25	выше нормы	
$k-1$	10	8	6	6	8	8	—	
q_i (в %%)	52	50	85	2	60	15	недостоверны	

Обнаружилось довольно неожиданно неплохое соответствие текстов Саллюстия и Ливия в частоте длин ЦП (см. табл. 8), чего, однако, нельзя сказать безоговорочно о текстах Саллюстия и Тацита. Следовательно, лаконичность в выражении, свойственная, по мнению исследователей, обоим этим авторам, была различного свойства. Результаты сравнения текстов Саллюстия и Марцеллина не приводятся, ибо уже отдельные значения χ^2 превышают допустимое для всей совокупности значение.

Таблица 9

Сравнение текста Ливия с текстами других авторов по частоте длин ЦП

Авторы	Лив.—Тац.		Лив.—Марц.	
Выборки	II—V	I—VII	VI—III	IV—V
χ^2	27.03	21.35	выше нормы	
$k-1$	10	9	—	
q_i (в %%)	<1	1	недостоверны	

Тексты Ливия, как и Саллюстия, вполне закономерно обнаружили свою неоднородность с текстами Тацита и Марцеллина (см. табл. 9).

Таблица 10

Сравнение текста Тацита с текстом Марцеллина

Выборки	III—II	IV—VI	VII—VII
χ^2	13.8	10.96	21.07
$k-1$	10	10	9
q_i (в %%)	18	37	<2.5

Как видно из таблицы 10, сравнение привело к противоположным результатам. Выборки из текстов Тацита и Марцеллина не обнаружили достаточно высокой степени однородности — одна пара выборок относится к зоне среднего соответствия, другая пара — к зоне никакого соответствия. Третья пара (обе выборки из последних книг произведений) обнаружила неоднородность.

Подводя итоги сравнения различных текстов относительно частоты длин ЦП приходим к выводу, что в пределах одного жанра — в данном случае латинской исторической прозы — нами получены противоречивые результаты, что вполне естественно, так как даже при сравнении выборок из одного текста получались недостоверные значения χ^2. Очевидно, что при переходе от выборок из одного текста к выборкам, представляющим собой различные тексты, уровень однородности падает.

Тем не менее метод сопоставления текстов различных авторов позволил сделать некоторые интересные выводы:

1) частота длин цельных предложений роднит текст Тацита с текстом Цезаря, текст Ливия — с текстом Саллюстия;

2) текст Марцеллина может оказаться однородным с текстами Цезаря или Тацита, но не Саллюстия и Ливия;

3) текст Ливия однороден только с текстом Саллюстия и ни с каким текстом другого автора;

4) текст Тацита может оказаться однородным с текстом Саллюстия.

Последним шагом эксперимента было сравнение результатов всех семи выборок из текста автора с аналогичным материалом из текста другого автора. Ввиду трудности подсчетов были сопоставлены тексты только тех авторов, для которых была выявлена однородность при сравнении отдельных выборок. Однако подобный эксперимент с большими массивами текстов Цезаря и Тацита дал недостоверный результат (при $k=15-1$ получено значение χ^2, равное 34.15, чему соответствует q_i менее 0.5%); зато тексты Саллюстия и Ливия обнаружили высокую степень однородности ($k=13-1$, $\chi^2=8.47$, $q_i=75\%$) во всей совокупности.

II. Однородность текстов на уровне ЭП

Методика исследования применена такая же, как при рассмотрении ЦП. Напомним, что *объем используемого материала* — семь выборок из произведений каждого автора длиною в 200 ЭП, всего около 1400 ЭП для одного автора.

Таблица 11

Количество ЭП в тексте Тацита

Классовый интервал	№№ классов	№№ выборок							Всего ЭП в каждом классе
		I	II	III	IV	V	VI	VII	
1—5	I	79	87	69	72	82	87	64	540
6—10	II	82	63	77	70	72	63	79	506
11—15	III	26	32	35	41	28	30	34	226
16—20	IV	13	10	12	14	13	15	12	89
21—25	V	2	6	7	3	7	5	7	37
26—30	VI	—	1	—	1	—	—	2	4
31—35	VII	—	1	—	1	—	—	1	3
36—40	VIII	—	—	—	—	—	—	0	0
41—45	IX	—	—	—	—	—	—	0	0
46—50	X	—	—	—	—	—	—	0	0
51—55	XI	—	—	—	—	—	—	1	1
Всего ЭП		202	200	200	202	202	200	200	1406

Проиллюстрируем способ обработки материала на отрывках из «Анналов» Тацита (см. табл. 11).

Необходимо отметить изменение интервала при разбивке ЭП на классы: он равен пяти словам вместо семи при разбивке на классы ЦП. Необходимость сокращения классового интервала вызвана тем, что ЭП в среднем короче ЦП в 2—3 раза (в зависимости от автора). Также упрощается и формула вычисления индивидуального *классового значения* χ^2, так как все выборки равны между собой по количеству ЭП:

$$\chi^2_i = \frac{(\mu_i - \nu_i)^2}{\mu_i + \nu_i}.$$

Суммировав индивидуальные классовые значения, получаем сразу истинное значение наблюденного χ^2. Результаты сравнения 21 пары выборок из «Анналов» Тацита представлены в таблице 12.

Таблица 12

Однородность частот ЭП в тексте Тацита

Выборки	$k-1$	χ^2	Оценка (в %%)
I—II	6	7.88	30
I—III	4	4.88	30
I—IV	6	6.83	35
I—V	4	3.55	50
I—VI	4	4.57	35
I—VII	10	9.51	50
II—III	6	5.86	45
II—IV	4	4.41	35
II—V	6	3.49	75
II—VI	6	3.12	80
II—VII	10	6.95	75
III—IV	6	4.61	60
III—V	4	2.11	70
III—VI	4	4.51	30
III—VII	10	4.22	92
IV—V	6	6.77	35
IV—VI	6	6.01	40
IV—VII	10	4.74	90
V—VI	4	1.28	85
V—VII	10	7.16	70
VI—VII	10	6.63	75

Распределяя выборки по зонам соответствия, устанавливаем, что в I зоне находится семь пар выборок, во II зоне — 5 пар выборок, в III — 9; в IV зоне не оказалось ни одной пары выборок. Эти результаты свидетельствуют о довольно высоком уровне стабильности текста Тацита относительно частот ЭП. Коренное отличие в распределении ЦП и ЭП по зонам однородности состоит именно в отсутствии неоднородных выборок и даже выборок

с низким соответствием на уровне ЭП. Суммарные же значения q_i, полученные при сложении результатов сравнения 21 пары выборок, почти тождественны у ЦП и ЭП — 1191% и 1187%.

Среди семи выборок из текста Тацита лучшее соответствие со всеми остальными обнаружила выборка из последней, 16-й, книги «Анналов» (совокупность q_i=452%). Худшее соответствие с остальными отрезками — у I выборки (кн. 1-я), у которой совокупность q_i составляет 230%. Аналогичным образом обстоит дело с выборкой I и на уровне ЦП, но совокупность q_i там еще ниже и равна 205%. Начальные и конечные части произведения нередко имеют свои особенности в языке и в стиле. Все пары выборок из текста Тацита, где одним из компонентов является 1-я книга, располагаются в одной зоне среднего соответствия или находятся на ее верхней границе.

Необходимо проанализировать вопрос о том, как изменяется однородность двух выборок при переходе с одного уровня на другой — из ЦП в ЭП и наоборот.

Возьмем *несколько пар выборок, однородных относительно* частот ЦП:

выборки	на уровне ЦП	на уровне ЭП	зоны
II—III	q_i = 99%	q_i = 45%	I → III
IV—V	q_i = 97%	q_i = 35%	I → III
II—VII	q_i = 95%	q_i = 75%	I → I/II

Замечаем, что все пары выборок при переходе на уровень ЭП переместились в более низкую зону однородности.

Рассмотрим, как ведут себя выборки, неоднородные на уровне ЦП, при переходе на уровень ЭП:

выборки	на уровне ЦП	на уровне ЭП	зоны
I—V	q_i = 9%	q_i = 50%	IV → II/III
I—VI	q_i = 1%	q_i = 35%	IV → III

Выборки, неоднородные на уровне ЦП, на уровне ЭП переместились в более высокую зону.

Будем исходить из уровня ЭП:

выборки	на уровне ЭП	на уровне ЦП	зоны
III—VII	q_i = 92%	q_i = 80%	I → I
IV—VII	q_i = 90%	q_i = 60%	I → II

Замечаем, что выборки, однородные на уровне ЭП, при переходе на уровень ЦП хотя и перестают быть однородными, но все-таки показывают довольно высокий уровень соответствия.

Как было сказано ранее, неоднородных выборок относительно частот ЭП у Тацита не оказалось.

Из всего вышеизложенного следует, что на материале текстов из «Анналов» Тацита не удалось обнаружить параллелизма с точки зрения однородности/неоднородности при рассмотрении одной и той же выборки на разных уровнях — ЦП и ЭП.

Остановимся на вопросе о соответствии частот отдельных классов ЭП. У Тацита насчитывается 5 полных классов. Наименьшее значение χ^2 (4.2) и, таким образом, лучшее соответствие по всем выборкам обнаружил IV класс с объемом ЭП в 16—20 слов. Наивысшее значение χ^2 (22.35) и худшее соответствие у I класса ЭП (объем 1—5 слов); этот класс является модальным.

Таков был круг проблем, рассмотренных на материале отрывков из «Анналов» Тацита. Однако для полноты картины необходимо привлечь результаты изучения подобных вопросов и у других авторов.

Итоги последовательного сопоставления 21 пары выборок на уровне ЭП в тексте Цезаря показали, что как в I зоне однородности (высокое соответствие), так и в III зоне (среднее соответствие) оказалось по 7 пар выборок, в зоне хорошего соответствия — 5 пар, в зоне низкого соответствия — 2 пары (IV—V, IV—VI), обе с недостоверным значением χ^2. Безусловно, частота длин ЭП обнаруживает у Цезаря бóльшую однородность, чем частота длин ЦП, если сравнить такие цифры: 1) в двух верхних *зонах на уровне ЭП располагаются 57% всех пар выборок*, на уровне ЦП — 38%; 2) в зоне низкого соответствия на уровне ЭП находятся 2 пары выборок, на уровне ЦП — 5 пар.

Однако самым любопытным оказалось выявление на уровне ЭП неоднородности отрезков из 4-й и 6-й книг. Неоднородность этих отрезков интуитивно ощущалась и ранее, но на уровне ЦП она не была обнаружена. Напомним, что эти книги имеют полярные коэффициенты комплексности ЦП: 3.7 — для 4-й книги и 2.35 — для 6-й. *Выяснилось, что такой разрыв в комплексности ЦП следует считать существенным: выборки оказались неоднородными на более низком уровне — на уровне ЭП.* Добавим, что выборка из 4-й книги вообще плохо согласуется со всеми остальными: совокупность q_i на уровне ЭП для этой книги составляет всего 112.5%; это самый низкий результат из засвидетельствованных для всех выборок у всех авторов.

А каково положение выборок из 1-й и 7-й книг? Первая отлично согласуется со всеми остальными (хуже всего — с выборкой из 4-й книги, но при этом q_i=0.4), как это было выявлено и на уровне ЦП. 7-я книга в области ЭП имеет лучшие соответствия с остальными книгами, чем в области ЦП.

Приведем совокупные данные q_i для 1-й, 4-й и 7-й книг:

Выборки из книг	I	IV	VII
q_i в ЦП (в %%)	325	175	135
q_i в ЭП (в %%)	420	112.5	345
всего %%	745	287.5	480

Не остается сомнения, что 4-я книга «Записок о Галльской войне» Цезаря по нескольким синтаксическим признакам — комплексности ЦП, частоте длин ЦП и ЭП — резко отличается от всех остальных книг произведения.

Переходя к вопросу об однородности текста Саллюстия на уровне ЭП, ограничимся сравнением распределения сопоставляемых попарно выборок по зонам однородности на уровнях ЦП и ЭП (см. табл. 13).

Таблица 13

Распределение выборок по зонам однородности на уровнях ЦП и ЭП в тексте Саллюстия

Зоны	I	II	III	IV
Уровень ЦП	7	7	6	1
Уровень ЭП	3	8	4	6

В тексте Саллюстия поражает попадание большого числа сопоставляемых на уровне ЭП выборок в зону низкого соответствия, тогда как у Тацита эта зона оказалась свободной, у Цезаря в нее попали всего 2 пары выборок. Из шести пар выборок IV зоны у Саллюстия три следует признать неоднородными, три оставшиеся имеют низкий уровень соответствия (q_i=0.2). Неудивительно, что понизилась и общая совокупность q_i для всех сопоставляемых выборок на уровне ЭП (905%) по сравнению с уровнем ЦП (1220%).

Большое количество неоднородных на уровне ЭП выборок у Саллюстия дает возможность проследить изменения, происходящие с этими выборками на уровне ЦП (чего мы не могли осуществить на материале «Анналов» Тацита):

выборки	на уровне ЭП	на уровне ЦП	зоны
I—III	$q_i = 10\%$	$q_i = 60\%$	IV → II
IV—V	$q_i = 10\%$	$q_i = 60\%$	IV → II
I—IV	$q_i = 5\%$	$q_i = 25\%$	IV → IV/III

Замечаем, что все выборки с недостоверными значениями q_i на уровне ЭП переместились на уровне ЦП в более высокие зоны однородности.

С другой стороны, три однородные между собой на уровне ЦП выборки — II, III и V — на уровне ЭП переместились в более низкие зоны:

выборки	на уровне ЦП	на уровне ЭП	зоны
II—III	$q_i = 95\%$	$q_i = 50\%$	I → II/III
II—V	$q_i = 95\%$	$q_i = 70\%$	I → II
III—V	$q_i = 95\%$	$q_i = 65\%$	I → II

Подтверждается вывод, сделанный первоначально на материале отрывков из «Анналов» Тацита, об отсутствии параллелизма

в поведении сопоставляемых выборок на двух разных уровнях — ЦП и ЭП — относительно однородности/неоднородности.

По вопросу о том, какая из выборок имеет лучшее соответствие со всеми остальными, а какая — худшее, материал «Югуртинской войны» Саллюстия также дал любопытные результаты. Оказалось, что начало произведения (гл. 5—17) хуже всех других отрезков согласуется с остальными выборками (совокупность q_i=160%). Именно этот отрезок оказался неоднородным с выборками III и IV. Отличие начала от прочих частей текста было отмечено и для «Анналов» Тацита. Но у Саллюстия особое положение начала произведения тем более показательно, что второй отрывок, следующий за первым без интервала (гл. 17—28), имеет лучшее соответствие со всеми остальными выборками (совокупность q_i=340%).

Из шести полных классов ЭП у Саллюстия лучшим соответствием по всем выборкам отличается II класс (объем 6—10 слов, совокупный χ^2=7.3); это класс модальный. Таким образом, еще раз подтверждается тезис о том, что стабильность частоты длин предложений нельзя непосредственно связывать с большой величиной частоты, которая в одних случаях оказывается устойчивой, в других, напротив, неустойчивой.

Обратимся к исследованию однородности текста на уровне ЭП в выборках из книг Ливия.

Таблица 14

Распределение выборок по зонам однородности на уровнях ЭП и ЦП в тексте Ливия

Зоны	I	II	III	IV
Уровень ЦП	9	6	3	3
Уровень ЭП	9	9	3	0

Данные таблицы 14 красноречиво свидетельствуют, во-первых, о перераспределении выборок по зонам однородности на уровне ЭП по сравнению с уровнем ЦП, во-вторых, о прекрасном соответствии между выборками из сочинения Ливия в отношении частоты ЭП разной длины. В двух верхних зонах располагается 86% сопоставляемых выборок, зона низкого соответствия пуста. Необычайно высокой, как ни у какого другого автора, оказалась совокупность q_i по всем выборкам на уровне ЭП — 1452% (в области ЦП соответствующая величина равна 1293%).

Вопрос о том, какая из выборок лучше других согласуется со всеми остальными, для Ливия не стоит так остро, как для других авторов. Разрывы в совокупностях q_i, соответствующих каждой выборке, настолько невелики, что имеет смысл выделить

только выборку V (кн. 33-я) как наиболее однородную со всеми выборками (совокупность q_i=475%).

Принимая во внимание большое количество однородных выборок в тексте Ливия на разных уровнях, было решено сделать еще одну попытку отыскать параллелизм в поведении выборок на уровне ЦП и уровне ЭП, отсутствие которого обнаружили ранее тексты Тацита, Цезаря и Саллюстия. Однако и на этот раз результат оказался отрицательным: параллелизма нет, как показывают следующие данные:

выборки	на уровне ЦП	на уровне ЭП	зоны
II—VI	q_i = 98%	q_i = 30%	I → III
I—VI	q_i = 95%	q_i = 60%	I → III
II—V	q_i = 90%	q_i = 85%	I → I
VI—VII	q_i = 90%	q_i = 75%	I → I/II

С некоторой оговоркой можно считать однородными на обоих уровнях только выборки II—V.

выборки	на уровне ЭП	на уровне ЦП	зоны
II—III	q_i = 97.5%	q_i = 55%	I → II
IV—V	q_i = 92%	q_i = 80%	I → I
I—V	q_i = 90%	q_i = 75%	I → I/II

Выборки, однородные на уровне ЭП, на уровне ЦП не имеют адекватных соответствий.

В шести полных классах ЭП лучшее соответствие зафиксировано в IV классе (совокупный χ^2=6.88), худшее в I классе (совокупный χ^2=20.54); частоты модального (II) класса отличаются неустойчивостью (совокупный χ^2=16.03).

И, наконец, проанализируем однородность текста на уровне ЭП в произведении самого позднего римского историографа Марцеллина.

Таблица 15

Распределение выборок по зонам однородности на уровнях ЭП и ЦП в тексте Марцеллина

Зоны	I	II	III	IV
Уровень ЦП	7	3	5	6
Уровень ЭП	4	5	8	4

Распределение выборок по зонам однородности у Марцеллина (см. табл. 15) как на уровне ЦП, так и на уровне ЭП очень невыразительно. В области ЭП преобладают выборки со средней степенью однородности — от 50 до 25%. Совокупность q_i по всем выборкам составляет 920%, что свидетельствует о более низкой степени однородности текста Марцеллина в области ЭП по сравнению с областью ЦП.

Крайние выборки — I и VII — противостоят друг другу по степени однородности с остальными выборками: выборка из начала произведения имеет лучшее соответствие (совокупность q_i = =360%), выборка из последней книги — худшее соответствие (совокупность q_i=140%). В сочетании с выборкой VII две другие выборки — II и V — дают на обоих уровнях столь низкое значение q_i, что данные выборки можно считать практически неоднородными как на уровне ЦП, так и на уровне ЭП. Таким образом, Марцеллин — единственный автор, у которого удалось обнаружить параллелизм в плане однородности выборок, сопоставляемых на разных уровнях членения предложений.

У Марцеллина полных классов оказалось более, чем у других авторов, а именно 8. Наибольшей однородностью по всем выборкам отличается II (модальный) класс с объемом ЭП в 6—10 слов (совокупный χ^2 равен 10.51). Худшее соответствие между выборками в IV и VI классах (совокупный χ^2 соответственно равен 27.6 и 27.17).

Далее необходимо перейти от наблюдений над текстом отдельного автора к обобщениям, касающимся однородности текста в области ЦП и ЭП у всех представителей римской исторической прозы, вместе взятых.

Таблица 16

Распределение выборок по зонам однородности на уровнях ЭП и ЦП в текстах латинских историографов

Зоны	I	II	III	IV
Уровень ЦП	31	31	24	19 (8)
Уровень ЭП	30	32	31	12 (7)
Всего	61	63	55	31

Как явствует из таблицы 16, в двух первых зонах заметных различий между уровнями ЦП и ЭП не наблюдается. В зоне среднего соответствия у ЭП лучший результат за счет сокращения числа выборок в зоне низкого соответствия. Числа в скобках указывают на количество неоднородных пар выборок, оно довольно значительно в обоих типах предложений. Напомним, что у двух авторов — Тацита и Ливия — зона низкого соответствия на уровне ЭП оказалась пустой, так что семь пар неоднородных выборок распределяются между Цезарем (2), Саллюстием (3) и Марцеллином (2).

На основании анализа таблицы 17 можно сделать вывод о том, текст какого из авторов следует считать наиболее однородным с точки зрения частоты длин и ЦП, и ЭП.

Таблица 17

Степень однородности текстов на синтаксическом уровне

Автор	Цез.	Салл.	Лив.	Тац.	Марц.	Всего
Совокупность q_i ЦП (в %%)	900	1220	1293	1191	1011	5615
Совокупность q_i ЭП (в %%)	1146.5	905	1452.5	1187	920	5611
Всего	2046.5	2125	2745.5	2378	1931	

Таким автором бесспорно является Ливий, текст которого, как мы видели, отличается высокой степенью однородности как на уровне ЦП, так и в особенности на уровне ЭП. За Ливием с немалым разрывом следуют Тацит, далее Саллюстий, Цезарь, и замыкает перечень Марцеллин. Любопытно, что совокупность q_i для всех авторов, вместе взятых, оказалась почти тождественной на уровне ЦП и на уровне ЭП, но неравномерно распределенной между авторами.

Анализ изменений, наблюдаемых в степени однородности выборок при переходе от одного уровня членения предложений к другому (от ЦП к ЭП и наоборот) также позволил сделать некоторые обобщения.

Исходя из уровня ЦП, мы можем констатировать, что выборки с высокой степенью однородности обычно перемещаются на уровне ЭП в более низкие зоны — во II—III; из 16 случаев, рассмотренных у всех авторов, такие переходы отмечены в 13 случаях; 3 пары выборок удержались в I зоне, причем две из них можно с некоторой натяжкой считать однородными на обоих уровнях. Переходов из I в IV зону не засвидетельствовано.

Подобным же образом обстоит дело и на уровне ЭП. Из 15 пар сопоставляемых выборок с высоким соответствием 12 переместились в более низкие зоны — II и III, одна пара выборок оказалась на грани между III и IV зонами, две пары выборок мы условно признали однородными.

Отсюда следует, что возможность оказаться однородными на обоих уровнях членения предложений — ЦП и ЭП — у выборок не так уж велика — всего 13—14%. Возможность того, что одна и та же пара выборок на одном уровне окажется однородной, а на другом неоднородной, практически равна нулю.

Исходя из уровня ЦП, мы можем констатировать, что выборки с низким (или недостоверным) значением χ^2, располагающиеся в IV зоне однородности, на уровне ЭП обычно перемещаются в более высокие зоны. Так, из 14 случаев, рассмотренных у всех авторов, в 9 случаях выборки переместились во II—III зоны, а одна пара выборок оказалась даже на грани между I и II зонами. Три пары выборок, будучи неоднородными на уровне ЦП, хотя и превысили на уровне ЭП допустимые значения q_i, однако остались в пределах IV зоны. И лишь одна пара выборок (из текста Саллюстия, I—IV), в противовес всем остальным, занимая положение между III и IV зонами, на уровне ЭП обнаружила неоднородность этих выборок.

Отсюда следуют, на наш взгляд, два немаловажных вывода:

1) возможность того, что выборки с низким уровнем однородности или даже неоднородные в области ЦП, окажутся на уровне ЭП в более высоких зонах, составляет 70%;

2) возможность того, что выборки с плохим соответствием на уровне ЦП окажутся неоднородными и на уровне ЭП, равна 7%, т. е. в 10 раз меньше.

Что касается выборок с низким соответствием q_i на уровне ЭП, то их может совсем не оказаться, как это отмечено для текстов Ливия и Тацита, или, если таковые есть, то на уровне ЦП они обычно перемещаются в более высокие зоны (6 случаев из 8). Две пары выборок (у Марцеллина) с чрезмерно низкими значениями χ^2 можно практически считать неоднородными.

Попытаемся также выяснить вопрос о том, какое соответствие имеют у наших авторов выборки из начала и конца произведений со всеми остальными выборками (см. табл. 18); цифры в этой таблице означают степень однородности: 1 — лучшее соответствие со всеми остальными выборками, 7 — худшее соответствие.

Таблица 18

Однородность выборок из начала и конца произведений на уровнях ЦП и ЭП

Автор	Цез.		Салл.		Лив.		Тац.		Марц.	
Уровень	ЦП	ЭП	ЦП	ЭП	ЦП	ЭП	ЦП	ЭП	ЦП	ЭП
Начало	2	1	4	7	3	2	7	7	4	1
Конец	7	5	6	4	6	7	3	1	7	7

Совершенно очевидно, что у Тацита начало «Анналов» противопоставлено всем остальным выборкам по обоим параметрам — как по частоте длин ЦП, так и по частоте длин ЭП. У Саллюстия начало произведения отличается от остальных его частей на уровне ЭП. Напротив, выборка из конца произведения противостоит всем остальным на обоих уровнях у Марцеллина, на уровне ЦП у Цезаря.

Итак, для большинства авторов правильным будет вывод о том, что начальная выборка лучше согласуется со всеми остальными выборками при том и другом членении текста, чем конец произ-

ведения, который, как правило, более резко отличается по своим параметрам от других частей произведения.

Что касается вопроса о поведении отдельных полных классов ЦП и ЭП, то здесь можно только повторить вывод, уже неоднократно полученный нами при рассмотрении материала разных авторов: устойчивость или неустойчивость частот отдельного класса нельзя прямо связывать с большой величиной частоты; самая высокая частота (в модальном или соседнем с ним классе) может быть как стабильной, так и нестабильной.

Переходим к рассмотрению проблемы об однородности между текстами разных авторов на уровне ЭП (см. табл. 19—21).

Таблица 19

Сравнение текста Цезаря с текстами других авторов на уровне ЭП

Авторы	1	Цез.—Салл.				Цез.—Лив.		
Выборки	2	I—I	IV—IV	VII—VII	VI—II	VII—IV	III—V	IV—II
χ^2	3	4.27	4.22	3.29	12.91	4.95	6.6	16.44
$k-1$	4	6	6	6	9	6	11	9
q_i (в %%)	5	65	65	75	18	55	82	7

Таблица 19 (*продолжение*)

Авторы	1	Цез.—Тац.			Цез.—Марц.	
Выборки	2	VII—III	III—VI	VI—II	VI—VI	VII—VII
χ^2	3	1.47	9.22	17.32	12.14	27.4
$k-1$	4	5	11	8	11	6
q_i (в %%)	5	92	60	<2.5	35	0

Наилучшее соответствие на уровне ЭП обнаружили тексты Цезаря и Саллюстия (см. табл. 19): из четырех сопоставлений три оказались в зоне хорошего соответствия и лишь одно сопоставление дало значение χ^2, приближающееся к критическому. Напомним, что на уровне ЦП сравниваемые тексты Цезаря и Саллюстия оказались неоднородными. Такое противоречие между двумя уровнями у названных авторов вполне объяснимо: частота длин ЭП у этих авторов сходна, но комплексность ЦП различна; отсюда и различие в частоте длин ЦП.

Сравнение текста Цезаря с текстами Ливия и Тацита привело к противоположным результатам, показывающим в большинстве случаев значительную степень однородности, но в некоторых —

неоднородность отрезков. С текстом Марцеллина текст Цезаря на обоих уровнях имеет мало общего.

Всевозможные комбинации выборок Саллюстия и Ливия каждый раз дают сходные результаты, свидетельствующие о высокой степени однородности (не ниже III зоны) текстов двух авторов (см. табл. 20). Напомним, что аналогичную картину мы наблю-

Таблица 20

Сравнение текста Саллюстия с текстами других авторов на уровне ЭП

Автор	Салл.—Лив.				Салл.—Тац.		Салл.—Марц.	
Выборки	I—I	III—IV	II—VII	IV—VI	VII—IV	I—VI	IV—IV	II—I
χ^2	2.92	6.16	7.62	12.56	4.37	14.07	30.52	12.22
$k-1$	8	8	9	13	6	5	11	9
q_i (в %%)	95	62	55	40	62	2	<0.5	20

дали и при сравнении выборок из произведений Саллюстия и Ливия на уровне ЦП. Следовательно, есть полное основание признать тексты этих историографов по двум параметрам — частоте длин ЦП и ЭП — в достаточной степени однородными. Впрочем, сравнение всего массива выборок из текстов Саллюстия и Ливия на уровне ЭП дало хотя и достоверный результат ($q_i=30\%$), но все же не столь высокий, как при членении текста на более крупные синтаксические единицы — ЦП, где было получено значение q_i, равное 75%.

Результаты сравнения текстов Саллюстия и Тацита, как и на уровне ЦП, носят противоречивый характер; с текстом Марцеллина текст Саллюстия, как и Цезаря, имеет мало общего (см. табл. 20).

Сравнение текста Ливия с текстами других авторов на уровне ЭП дало результаты, аналогичные результатам сравнения текста Саллюстия с текстами Тацита и Марцеллина (см. табл. 21).

Таблица 21

Сравнение текста Ливия с текстами других авторов на уровне ЭП

Автор	Лив.—Тац.		Лив.—Марц.		
Выборки	III—II	V—IV	IV—VII	II—V	VII—I
χ^2	12.98	5.44	10.72	13.85	15.7
$k-1$	6	8	6	10	9
q_i (в %%)	5	70	10	15	7

Сопоставление выборок из произведений Тацита и Марцеллина дало преимущественно недостоверные значения χ^2.

Итак, единственные авторы, чьи тексты бесспорно обнаружили довольно значительную степень однородности на обоих рассматриваемых уровнях — это Саллюстий и Ливий. Тексты Цезаря и Тацита могут дать результаты прямо противоположного характера, свидетельствующие как об однородности выборок, так и о неоднородности.

Очень противоречивые итоги дает сравнение текстов Цезаря и Саллюстия, Цезаря и Ливия: значительная степень однородности на уровне ЭП сочетается с неоднородностью текстов на уровне ЦП.

Сравнение текстов самых лаконичных римских историографов Саллюстия и Тацита показывает, что частота длин как ЦП, так и ЭП у этих авторов может иметь хорошее соответствие, а может и обнаружить существенные различия.

Наименее однородными на обоих рассматриваемых уровнях следует считать тексты Цезаря и Марцеллина, еще менее — Ливия и Тацита, Тацита и Марцеллина. Тексты Саллюстия и Марцеллина, Ливия и Марцеллина практически следует считать неоднородными.

При переходе от текста одного автора к текстам разных авторов вероятность того, что две выборки окажутся неоднородными на обоих уровнях, значительно возрастает. Если в пределах одного текста такая вероятность, как мы установили, равнялась всего 7%, то при сравнении разных текстов она достигает $\approx$50%.[2] Отмечено также два случая, когда выборки, неоднородные на одном уровне, на другом уровне практически должны рассматриваться как однородные (q_i=85%); в пределах же одного текста и при сквозном сопоставлении выборок подобные случаи не встретились. Следовательно, при переходе от одного к разным текстам вероятность того, что одна и та же пара выборок на разных уровнях членения текста может оказаться однородной/неоднородной, возрастает от нуля до 10%.

III. Однородность текста в двух произведениях одного автора

В заключение статьи коснемся еще одного немаловажного вопроса: об однородности текста в разных произведениях одного автора (в данном случае материалом послужили «Югуртинская война» и «Заговор Катилины» Саллюстия).

Небольшое произведение «Заговор Катилины» было расписано почти полностью (за исключением вставных речей) и дало три выборки объемом 200 ЭП каждая. Следовательно, эти три выборки в соответствии с композиционным членением можно рассматривать как начало (I), середину (II) и конец произведения (III).

Прежде всего выясним, однородны ли между собой отдельные части «Заговора Катилины» («ЗК») на уровне ЭП. Результаты сопоставления даны в таблице 22.

Таблица 22

Однородность выборок из «ЗК» на уровне ЭП

Выборки	χ^2	$k-1$	q_i (в %%)
I—II	8.84	7	25
I—III	3.3	6	75
II—III	14.63	7	<5

Начало и конец произведения обнаружили довольно высокую степень соответствия (q_i=75%). Центральная часть отличается по частоте длин ЭП от начала произведения (q_i=25%) и неоднородна с концом ($q_i < 5\%$). Если заглянуть в результаты сравнения выборок из «Югуртинской войны», то окажется, что и там центральная часть произведения (выборка IV) неоднородна с начальной выборкой и плохо согласуется с конечной (q_i=20%).

Теперь перейдем непосредственно к выяснению вопроса о том, однородны ли между собой соответствующие части «Заговора Катилины» и «Югуртинской войны» Саллюстия.

Таблица 23

Однородность композиционных частей двух произведений Саллюстия на уровне ЭП

Выборки из разных частей произведений	χ^2	$k-1$	q_i (в %%)
Начало	3.84	6	70
Середина	5.84	9	75
Конец	3.21	6	80

Из таблицы 23 явствует, что соответствующие части двух произведений одного автора на уровне ЭП имеют очень неплохие соответствия. Отсюда следует, что композиционное членение накладывает определенный отпечаток на авторский стиль, так что некоторые параметры, в данном случае частоты длин ЭП, обнаруживают большое сходство в соответствующих частях разных по содержанию произведений. Ибо, если сравнить, например, центральную часть «Заговора Катилины» порознь с началом и концом «Югуртинской войны», то χ^2 регист-

[2] Из 19 пар выборок, неоднородных на уровне ЦП, 10 пар оказались неоднородными и на уровне ЭП.

рирует неоднородность этих композиционно не соответствующих друг другу частей.

Далее установим степень однородности между собой трех частей «Заговора Катилины» на уровне ЦП (см. табл. 24).

Таблица 24

Однородность выборок из «ЗК» на уровне ЦП

Выборки	χ^2	$k-1$	q_i (в %%)
I—II	11.74	9	25
I—III	6.57	9	70
II—III	10.3	9	30

На уровне ЦП наблюдается определенное сходство с положением дел на уровне ЭП: хорошее соответствие начальной и конечной выборок сочетается с низким соответствием начала и середины, конца и середины произведений.

Далее необходимо выяснить, однородны ли между собой соответствующие части «Югуртинской войны» и «Заговора Катилины».

Оказалось, что соответствующие части произведений Саллюстия, продемонстрировавшие высокое соответствие на уровне ЭП, при членении текста на более крупные синтаксические единицы — ЦП — или становятся неоднородными (при сравнении как начала произведений, так и конца), или обнаруживают лишь среднюю степень однородности (центральные части) (см. табл. 25).

Таблица 25

Однородность композиционных частей двух произведений Саллюстия на уровне ЦП

Выборки из разных частей произведений	χ^2	$k-1$	q_i (в %%)
Начало	21.36	8	<1.0
Середина	8.22	9	50.0
Конец	22.68	9	<1.0

Наше исследование показало, что композиционно соответствующие друг другу части двух произведений одного автора демонстрируют довольно значительную степень однородности относительно частоты длин ЭП. На уровне ЦП, где вступает в силу фактор комплексности предложений, однородность текстов или снижается, или переходит в неоднородность.

Л. В. Малаховский

ПРИНЦИПЫ ЧАСТОТНОЙ СТРАТИФИКАЦИИ СЛОВАРНОГО СОСТАВА ЯЗЫКА

Употребительность слова, измеряемая его частотой или рангом, является одной из его важнейших языковых характеристик. В работах по лексикологии и стилистике, по автоматическому анализу текста и лингвостатистике, по составлению словарей и совершенствованию методики преподавания языка, а также в психолингвистических исследованиях экспериментального характера нередко возникает необходимость в противопоставлении слов более частых более редким, в выделении в словарном составе языка нескольких частотных слоев (см., напр.: Семенова, 1975). Одним из первых необходимость частотной стратификации лексики осознал выдающийся американский психолог и лексикограф Э. Л. Торндайк, который в своем частотном словаре, предназначенном для учителей английского языка (Thorndike, 1921), разделил приводимые слова на классы, расположив их в порядке нарастания ранга слов («1-я сотня», «2-я сотня», . . . «1-я тысяча», «2-я тысяча» и т. д.). Подобный подход был применен и в ряде последующих частотных словарей английского языка (Horn, 1926; Rinsland, 1945).

Однако способ группировки слов, используемый в этих словарях, не всегда оказывался пригодным для решения конкретных исследовательских задач. Поэтому во многих работах предлагаются иные способы частотной стратификации лексики. Некоторые исследователи (Peters, 1936; Тышлер, 1966) кладут в основу стратификации ранг слова и выделяют классы, аналогичные тем, которые были предложены Торндайком в его первом частотном словаре. В большинстве работ, однако, разбиение словарного состава на классы производится на основе частоты слова. При этом разными авторами эта задача решается по-разному — как в отношении охвата диапазона частот, так и в смысле количества вы-

деляемых частотных участков и способа проведения границ между ними.

Так, в работе Д. Энтуисла, исследовавшего образование словесных ассоциаций у детей дошкольного возраста, охватываются слова с частотой от 25 миллионных и выше [1] (Entwisle, 1966), а в работе Г. Питерса, изучавшего зависимость между употребляемостью слов и их запоминаемостью — слова с частотой выше 30 миллионных (Peters, 1936). В исследованиях Дж. Холла и Г. Самби, также посвященных вопросу о запоминаемости слова, и в работе Р. М. Фрумкиной, сравнивающей объективные и субъективные оценки вероятностей слов, используются ограниченные диапазоны частот — от 1 до 100 [2] (Hall, 1954), от 1 до 1000 (Sumby, 1963) и от 70 до 7000 (Фрумкина, 1966). В психолингвистических экспериментах Д. Хауэса и Р. Соломона (Howes, Solomon, 1951) и в ряде других работ охватывается весь диапазон частот, включая нулевую (слова, отсутствующие в частотном словаре).

В не меньшей мере колеблется и количество выделяемых частотных участков — от 3—4 в работах Д. Энтуисла, Дж. Холла и др., до 14—15 в работах Р. М. Фрумкиной и Д. Хауэса — Р. Соломона. Но наибольший разнобой наблюдается в установлении границ между участками и соответственно в размерах участков. В большинстве случаев авторы проводят границы между участками произвольно, не опираясь ни на какие лингвистические или стилистические закономерности (см. табл. 1).

Таблица 1

Границы частотных участков, выделяемых разными авторами

Авторы	Количество участков	Частотные характеристики выделяемых участков
Энтуисл	3	>1000; $1000 \div 500$; $500 \div 25$
Холл	4	$100 \div 50$; 30; 10; 1
Самби	4	$1100 \div 900$; $110 \div 90$; $11 \div 9$; $1 \div 0$
PGC *	6	$\geqslant 100$; $99 \div 50$; $49 \div 30$; $29 \div 10$; $9 \div 0.22$; 0

* Phoneme-grapheme correspondences, 1966.

Как можно видеть из таблицы, границы между участками проводятся в каждом случае по-разному и сами участки оказываются не равными между собой по охвату частот: расстояния между

[1] Во всех упоминаемых здесь работах оценка частоты слов производилась по частотному словарю Торндайка-Лорджа (Thorndike, Lorge, 1944).

[2] Здесь и далее частоты указываются в миллионных.

центральными частотами участков колеблются даже у одного и того же автора от 5 до 500 и более.

Однако еще в 1951 г. Д. Хауэс и Р. Соломон положили в основу выделения частотных участков четкий принцип — равное логарифмическое расстояние между центральными частотами. Весь диапазон частот был разделен на 15 участков, центральные частоты которых отстояли друг от друга на одно и то же логарифмическое число ($\lg=0.3000$). К сожалению, в дальнейшем по этому пути пошли лишь немногие исследователи. Из авторов, упомянутых нами ранее, это Г. Самби, в работе которого используется логарифмическое расстояние 1.0000, и Р. М. Фрумкина, применяющая расстояние, равное 0.2000. Однако обе эти работы охватывают весьма ограниченный диапазон частот и имеют (так же, впрочем, как и работа Д. Хауэса и Р. Соломона) весьма конкретную практическую направленность.

Отсутствие общепринятого рационального способа частотной стратификации лексики делает результаты работ разных авторов несравнимыми между собой, что тормозит развитие исследований в различных областях лингвистики, связанных с учетом частотной характеристики слова, и прежде всего в области самой лингвостатистики. В связи с этим в настоящей работе делается попытка обосновать принципы, которые могли бы быть положены в основу выделения частотных слоев, или зон, в словарном составе языка.

Прежде всего, разбиение лексики на частотные зоны должно производиться не на основе ранга слова, а на основе его частоты. Частота слова, являющаяся мерой его повторяемости в текстах, имеет более непосредственное отношение к процессам запоминания и узнавания, кодирования и декодирования слова,[3] чем его ранг. Конечно, определенная зависимость между «стоимостью» декодирования слова и рангом, на которую указывает Б. Мандельбро (Mandelbrot, 1954, с. 16—17), также существует, но это зависимость опосредованная, поскольку ранг слова является величиной вторичной, определяемой на основе частоты. Зависимость же между процессами кодирования слова и частотой является прямой и непосредственной.[4] Неудивительно, что начиная с 40-х гг. буквально все исследователи (за исключением И. С. Тышлера) кладут в основу стратификации лексики частоту, а не ранг; отказался от рангового разбиения словарного состава и Э. Л. Торндайк в своем третьем частотном словаре (Thorndike, Lorge, 1944).

[3] Ср. в связи с этим высказывание Д. Хауэса о том, что «концепция частоты слова связана через процесс выбора слова с центральными проблемами теории языка» (Howes, 1964, с. 53).

[4] Об этом говорит и сам Мандельбро, отмечающий, что «. . .где-то существует кодировка, в которой стоимость одного слова пропорциональна одновременно частоте слова и его рангу» (Mandelbrot, 1954, с. 22).

Полностью неприемлемым является способ, при котором словарный состав разбивается на участки, равные по числу слов («1-я тысяча», «2-я тысяча» и т. д.), как это сделано, в частности, в работе И. С. Тышлера (Тышлер, 1966). Совершенно очевидно, что по своим вероятностным характеристикам начальное и конечное слова одного и того же высокочастотного участка (например, 1-й тысячи) гораздо сильнее различаются между собой, чем такие же слова низкочастотного участка (например, 20-й тысячи).

Разбиение на участки, пропорциональные логарифмам ранговых чисел (1—10 слово, 11—100 слово и т. д.), устранило бы этот недостаток, однако практически это осуществимо только в высокочастотной области словаря. В средне- и низкочастотных областях, где одну и ту же частоту имеют десятки, сотни и даже тысячи слов, пришлось бы равночастотные слова искусственно разносить по разным участкам или же создавать крайне неравномерные участки, которые к тому же оказались бы различными в разных словарях.

Далее, разбиение словарного состава на частотные зоны, должно опираться *не на абсолютные, а на относительные частоты*. Использование абсолютных частот, широко практикующееся в настоящее время, делает невозможным сопоставление данных, полученных на основе частотных словарей разного объема, а иногда приводит исследователей к неверным выводам. Так, в работе К. Б. Бектаева утверждается, что с увеличением объема словаря процент редких слов в нем снижается (Бектаев, 1978, с. 52). Это утверждение противоречит интуитивному представлению о том, что в больших словарях словник возрастает как раз за счет редких слов. Противоречие это вызвано тем, что, сравнивая долю редких слов в словарях разного объема, автор пользуется не относительными, а абсолютными частотами. Поскольку «редкие» слова определяются им как слова с абсолютной частотой 1 и 2 (независимо от объема словаря), то оказывается, что в словарях меньшего объема «редкие» слова имеют значительно бóльшую относительную частоту, чем в больших словарях. Неудивительно, что доля таких слов в малых словарях оказывается выше, чем в больших. Использование относительных частот позволит при выделении частотных зон непосредственно, без каких-либо преобразований, сравнивать данные разных частотных словарей независимо от их объема.

В качестве эталонной границы относительной частоты удобно использовать величину, ставшую уже традиционной в лингвостатистических исследованиях по английскому языку — одну миллионную. Использование такой единицы дает возможность выражать относительные частоты для подавляющего большинства слов целыми числами, преимущественно однозначными или двузначными.

Разграничение между частотными зонами должно осуществляться на основе *не арифметического, а логарифмического масштаба*. С одной стороны, диапазон частот лексических единиц в языке настолько велик (самое употребительное слово встречается в миллионы раз чаще, чем самые редкие слова), что в арифметическом масштабе он практически необозрим. С другой стороны, как показали исследования ряда авторов (Howes, Solomon, 1951, с. 53; Miller, 1954, с. 413 и др.), время, необходимое для осуществления в речевых механизмах человека процессов кодирования и декодирования слова, находится в логарифмической зависимости от его частоты. Поэтому частотные зоны, выделенные на основе логарифмического масштаба, будут более адекватно соответствовать тем вероятностным градациям, которым подчиняются аналоги частоты слова, имеющиеся в речевых механизмах (Фрумкина, 1966, с. 90). Как уже говорилось выше, некоторыми авторами логарифмический масштаб при выделении частотных участков лексики уже использовался, и притом весьма успешно; однако наличие большого числа работ, в которых применяется арифметический принцип, говорит о том, что необходимость логарифмического подхода исследователями в данной области еще недостаточно осознана.

Какова же должна быть величина частотной зоны, какой диапазон частот она должна охватывать?

По ряду соображений наиболее целесообразным представляется такое разграничение частотных зон, при котором каждая зона охватывала бы *диапазон, равный одному порядку*, т. е. при котором верхняя граница зоны характеризуется частотой в 10 раз большей, чем нижняя.

Практическое удобство такого подхода неоспоримо. Если в работе Д. Хауэса и Р. Соломона центральные частоты двух соседних зон отличались друг от друга на величину $\lg=0.3000$, а в работе Р. М. Фрумкиной на величину $\lg=0.2000$, что весьма неудобно для быстрых расчетов и сопоставлений, то при предлагаемом подходе они будут различаться между собой равно в 10 раз.

Кроме того, нельзя не признать, что частотная характеристика слова, заложенная в речевых механизмах человека, является весьма грубой и приблизительной. Интуитивно мы сравнительно легко делим слова языка на частые, средние по употребительности и редкие; возможно, по-видимому, выделить также слова «очень частые» и «очень редкие», что дает всего 5 градаций, но вряд ли наше языковое сознание способно на более тонкое членение. Это подтверждается и опытами ряда исследователей, в частности Холла (Hall, 1954), который показал, что различия в запоминаемости слов разной частоты выявляются, когда слова отстоят друг от друга по частоте на 1 порядок ($f=1$ и $f=10$), но не обнаруживаются, если частоты слов различаются по полпорядка и менее ($f=30$ и $f=50\div100$). Если выделять частотные зоны с интервалом в 1 порядок, то общее число зон, получаемых на материале самого большого из существующих частотных словарей английского

языка, основанного на текстах объемом в 18 миллионов словоупотреблений (Thorndike, Lorge, 1944), окажется равным 6, что близко к числу градаций, различаемых интуитивно.

Выделенные в соответствии с изложенными принципами частотные зоны [5] обозначаются нами, в порядке убывания частоты, буквами A, B, C, D, E и F. Количественные и качественные характеристики зон представлены в таблице 2.

Таблица 2

Количественные и качественные характеристики частотных зон

Частотная зона	Диапазон относительных частот (в миллионных)	Диапазон логарифмов относительных частот	Количество слов в зоне (по словарю Торндайка-Лорджа)	Количественная характеристика зоны	
				служебные слова	знаменательные слова
A	99 999—10 000	5.0—4.0	~10	частые	—
B	9999—1000	4.0—3.0	~100	средние	очень частые
C	999—100	3.0—2.0	~1 000	редкие	частые
D	99—10	2.0—1.0	~6 000	—	средние
E	9—1	1.0—0.0	~13 000	—	редкие
F	0.9—0.1	0.0—1.0	~20 000	—	очень редкие

Примечания. 1. Для большей наглядности логарифм верхней границы каждой частотной зоны, весьма близкий к целому числу, округлен до этого целого (например, в зоне A — 4.9999 до 5.0 и т. д.).

2. К зоне F нами отнесены также слова, встретившиеся в текстах 1 раз на 18 миллионов словоупотреблений, т. е. имеющие относительную частоту 0.05 и, следовательно, открывающие уже новую частотную зону (G). Эти слова, так же как и слова зоны F с абсолютной частотой 2 и 3 на 18 000 000 (относительные частоты 0.11 и 0.17), в опубликованные списки не включены: из 40 000 разных слов, встретившихся в текстах Торндайка-Лорджа, в частотные списки вошло 30 000 (с абсолютной частотой 4 и выше).

При таком разбиении в зону A попадают все слова, частоты которых выражены пятизначными цифрами, в зону B — четырехзначными и т. д., а в зону F — десятыми долями миллионной. Пользуясь данными словаря Торндайка-Лорджа, можно установить, что в английском языке в зону A входит около десятка слов, в зону B — около сотни, в зону C — около тысячи. В дальнейшем, однако, эта логарифмическая зависимость нарушается, и зоны D, E, F содержат уже по 6, 13 и 20 тысяч слов соответственно (см. также примечание 2 к табл. 2).

Изучение лексического наполнения частотных зон позволяет охарактеризовать зоны A и B как зоны преимущественно служебных слов, а остальные зоны — как зоны слов знаменательных.

[5] Впервые эти принципы были кратко сформулированы нами в работе: Малаховский, 1968, с. 30—35.

В зону A попадают самые употребительные служебные слова (the, of, in, to, a и др.), в зону B — такие служебные слова, как but, by, be, before, а в зону C — сравнительно небольшое число малоупотребительных служебных слов типа among или beyond. Что касается слов знаменательных, то в зоне A их нет совсем, в зоне B встречаются лишь немногие, очень употребительные (back, bring). Зону C можно охарактеризовать как зону частых знаменательных слов (child, city, car, cost), зону D — как зону слов средних по употребительности (duck, delicate, despair, dependent), зону E — как зону редких (explode, eyelid, evaporate, embankment), а зону F — как зону очень редких слов (falsetto, felonious, flotsam, faugh).

В исследовательской работе может возникнуть необходимость в более тонком различении частотных характеристик. В связи с этим предусмотрено разбиение каждой из частотных зон на два участка с логарифмическим интервалом между границами участков, равным 0.5. В этом случае частоты двух соседних участков будут различаться между собой в среднем на полпорядка, т. е. на величину, равную $\sqrt{10}=3.162$. Тогда зона E, например, будет состоять из участков с частотой 1÷3 и 4÷9, зона D — из участков с частотой 10÷31 и 32÷99 и т. д.

На основе выделенных таким образом частотных зон и участков были исследованы на материале английского языка распределение омонимов и омоформ по частоте, зависимость объема омонимического ряда от его частоты, зависимость объема конверсионного гнезда от частоты, зависимость между длиной омоформы и ее частотой и т. п. (см. нашу вторую статью в настоящем сборнике).

Предложенные принципы частотной стратификации лексики могут оказаться полезными для совершенствования и унификации лингвостатистической методики и способствовать развитию исследований в различных областях языкознания, связанных с использованием концепции частоты слова.

Часть III
АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТА

Е. М. Лукьянова

ИНФОРМАЦИОННАЯ БАЗА АВТОМАТИЧЕСКИХ СЛОВАРЕЙ

§ 1. Лингвистическое обеспечение автоматизированных систем переработки информации и банк данных

Сложные автоматизированные системы переработки информации, которые включают автоматизированные системы управления (АСУ), автоматизированные системы обработки данных (АСОД) и автоматизированные информационные системы (АИС), широко применяются в настоящее время при управлении крупными предприятиями, учреждениями, отраслями народного хозяйства и народным хозяйством в целом, а также при проведении научных и социологических исследований и создании систем искусственного интеллекта, так называемых лингвистических автоматов. При этом возникает необходимость автоматизации функций информационного обеспечения (ИО) задач пользователя этих систем в условиях «удобной вычислительной обстановки» (Информационные системы общего назначения, 1975, с. 5). Целям создания такой обстановки служат выделение процессов, связанных с техническим, информационным, математическим и организационным обеспечением решения задач пользователя для данной вычислительной системы, и последующая автоматизация выполнения этих процессов в виде автономных процедур, программно или аппаратурно встраиваемых в вычислительную систему в виде ее подсистем.

В плане технического и математического обеспечения эти процедуры *были предусмотрены уже при создании стандартного* hardware и software вычислительных систем третьего поколения, где, в частности, предусмотрены контроль и управление внешним оборудованием, создание физической и логической систем управления вводом-выводом (СУВВ), управление ресурсами системы, организация вычислительного процесса (управление задачами, заданиями и данными) и т. п. Что же касается системы автоматизации информационного обеспечения, то она находится сегодня либо на стадии проектирования и разработки, либо, в лучшем случае, на стадии освоения. Между тем необходимость автоматизации функции информационного обеспечения диктуется в первую очередь резким возрастанием мощности потока перерабатываемой информации и усложнением ее структуры в АСУ и АИС. Для решения поставленной проблемы имеются следующие предпосылки, способствующие созданию систем автоматизации функций ИО:

1) создана мощная техническая база, представленная вычислительными комплексами ЭВМ третьего поколения и, в частности, ЕС ЭВМ;

2) при решении широкого класса задач управления многократно используются данные, которые составляют единую информационную базу той или иной подсистемы или даже системы в целом;

3) определен основной принцип автоматизации процессов информационного обеспечения: организационное и функциональное отделение сбора, хранения и обновления информации от конкретных процессов выполнения задач.

Последний фактор и лег в основу концепции автоматизированного банка данных (БД), представляющего собой совокупность данных и организационных, технических, программных, языковых средств — совокупность, предназначенную для централизованного и интегрированного накопления и хранения данных и их многократного использования. Основными элементами БД являются информационная база данных (ИБ) на устройствах прямого доступа и система управления базой данных (СУБД), представляющая собой комплекс организационных и программных средств, обеспечивающих создание ИБ, хранение, обновление и поиск информации в ИБ.

В настоящее время в нашей стране и за рубежом принят следующий подход в разработке БД, в части создания СУБД. Технология разработки банков данных предусматривает:

1) ориентацию на серийно выпускаемые средства вычислительной техники;

2) базирование на принятую для операционных систем (ОС) этих ЭВМ организацию вычислительного процесса;

3) использование средств стандартного математического обеспечения для сопряжения пользователя с ЭВМ.

Средства организации файлов, предоставляемые стандартным техническим и математическим обеспечением вычислительных систем, отличаются статичностью, автономностью, жесткой привязкой к конкретной задаче. Они порождают многократное дублирование данных, ведущее к перерасходу памяти и увеличению стоимости хранения и обслуживания данных, а также неоправданно частую реорганизацию файлов.

В банках данных эти недостатки отсутствуют. Их основная функция состоит, во-первых, в обеспечении коллективного доступа

к информации со стороны пользователей, задач и других подсистем и, во-вторых, в поддержании в автоматизированной системе динамической информационной модели сложного управляемого или исследуемого объекта. Последнее реализуется путем запоминания в БД логических связей между отдельными типами данных и автоматизацией процесса их использования. Естественно поэтому, что банки данных становятся неотъемлемыми компонентами АСУ, АСОД и АИС, разрабатываемых для решения совокупности задач, использующих данные сложной структуры и большого объема. Такими системами в первую очередь являются проектируемые в настоящее время информационно-вычислительные системы общегосударственного значения: Автоматизированная система государственной статистики (АСГС), Автоматизированная система плановых расчетов (АСПР), Государственная система научно-технической информации (ГОСНТИ).

1.1. Лингвистическое обеспечение как часть информационного обеспечения АСУ и АИС

Важной проблемой, возникающей при создании систем автоматизации функций информационного обеспечения, является задача разработки развитых средств, необходимых для переработки неформализованной информации.

Это объясняется тем, что поток информации, идущий в АСУ и АИС, включает не только доступные для машинной обработки сведения — цифры, формулы, кодовые обозначения, но и большие массивы текста, написанного на естественном языке (ЕЯ) — текста, в котором информация представлена в неформализованном виде, в связи с чем возникает проблема обеспечения автоматического распознавания смыслового содержания документа и формализации этого смысла для дальнейшей обработки на ЭВМ. Известно, что такую задачу часто рассматривают как периферийную, имеющую слабую связь с дальнейшими задачами автоматизации информационного обеспечения. Эта точка зрения отражает еще бытующее до сих пор неверное представление о том, что весь процесс управления и переработки информации можно свести к математической формализации, полностью исключив из него участие человека. Между тем сегодня, несмотря на то что быстродействие серийных ЭВМ постоянно растет (увеличилось в сотни раз по сравнению с предыдущим десятилетием), а методы переработки информации внутри машин стремительно совершенствуются, приемы и средства формализации неформализованной информации остаются такими же, как 10—15 лет назад. Такая работа по-прежнему в большинстве вычислительных центров выполняется вручную, требует больших затрат времени и средств. Результаты ее во многом зависят от квалификации, опыта, внимания и личных качеств специалистов. В итоге между информационным потоком деловой документации и ЭВМ возникает барьер домашинной подготовки документов. Чтобы преодолеть его, необходимо разработать систему средств автоматического распознавания в неформализованном образе документа его инварианта в форме, доступной для обработки на ЭВМ. Эти средства должны быть организованы как часть информационного обеспечения автоматизированных систем, которую будем именовать лингвистическим обеспечением (ЛО).

С помощью средств лингвистического обеспечения в АСУ и АИС должны осуществляться следующие виды переработки текстов, составленных на естественных языках:

1) ввод управленческой документации в ЭВМ и вывод информации для принятия решений в задачах АСУ на ЕЯ;

2) атрибуция, аннотирование и реферирование документов, составленных на русском языке;

3) атрибуция, аннотирование и реферирование иностранных документов средствами русского языка;

4) машинный перевод иностранных текстов (МП);

5) поиск библиографической информации на русском и иностранном языках;

6) поиск фактографической информации в русском тексте;

7) диалог «человек—машина» на русском языке для задач 1—6;

8) различные лексикографические, грамматические, лингвостатистические исследования.

Перечисленные задачи объединим в класс лингвистических задач (класс ЛЗ), относительно которого далее будут рассматриваться все аспекты реализации системы ЛО и ее компонентов.

Создание системы лингвистического обеспечения связано с трудностями двоякого характера: во-первых, с формализацией неформализованной текстовой информации, во-вторых, с самой машинной реализацией ЛО.

I. Рассмотрим лингвистическую сторону проблемы.

1. Естественный язык представляет собой открытую систему передачи информации, опирающуюся на нечеткие множества лингвистических элементов. В память же ЭВМ может быть введено только закрытое, сокращенное описание ЕЯ (так называемый базовый язык, состоящий только из четких множеств лингвистических элементов). Это объясняется техническими возможностями ЭВМ (ограниченным объемом памяти), неразработанностью алгоритмов ассоциативного мышления и вынуждает нас задавать машине фиксированное (закрытое) описание открытой нефиксированной системы естественного языка.

2. Порожденный в результате взаимодействия смысловой информации, с одной стороны, и системы и нормы языка, с другой, текст обладает статистической структурой. Эта структура включает как элементы, общие для всех языков мира (средняя длина слова, уровень избыточности — так называемые универсалии), так и черты, специфические для определенного языка, стиля, подъязыка, индивидуальной манеры (колеблющиеся частоты отдельных слов,

словосочетаний, грамматических форм и конструкций), т. е. ЕЯ функционирует в условиях постоянно действующей антиномии *язык — идиолект*, в то время как для *базового языка* его *коллективные* и индивидуальные, межъязыковые и внутриязыковые интерпретации совпадают.

Эффективное функционирование системы «человек—машина—человек» предполагает, что в ЭВМ вводится в виде базового языка информационная модель описания понятий и объектов внешнего мира, так называемая воспроизводящая инженерно-лингвистическая модель. Эта модель должна быть согласована с теми представлениями внешнего мира, которыми оперируют человек — источник и передатчик информации и человек — приемник информации, что практически невозможно, так как эти представления *не совпадают, а степень согласования может лишь приближаться* к реальной.

Отсюда следует общий вывод, что машинная переработка текста принципиально не может быть стопроцентной. При разработке лингвистического обеспечения необходимо ориентироваться на такой уровень правильной автоматической переработки текстов, при котором живой приемник сообщения, снова пользующийся полным и открытым естественным языком, а также располагающий достаточными сведениями о предмете сообщения, быстро и удовлетворяющим его образом восстанавливает, трактуя в целом правильно, исходное содержание текста.

II. Вопросы машинной реализации системы лингвистического обеспечения в настоящее время также являются неразрешенными. Большое количество отдельных программ реализации лингвистических алгоритмов не объединено в стройную систему. Имеющиеся системы автоматической переработки текста не используют возможностей ЭВМ третьего поколения. Единая информационная база, необходимая для решения лингвистических задач, отсутствует вообще.

Естественно, что разработка ЛО как системы требует системного подхода. Это означает, что все аспекты разработки должны быть увязаны между собой и при рассмотрении каждого вопроса *необходимо учитывать влияние* принимаемых решений на характеристики системы в целом. Разработка системы должна быть многоуровневой; на каждом уровне группы соответствующих специалистов должны решать свои специфические задачи. Предлагаются следующие уровни разработки системы: первый уровень — общесистемный, второй уровень — информационный, третий уровень — разработки технических средств, четвертый уровень — разработки математического обучения. Одним из основных вопросов, решаемых при системном подходе к разработке ЛО, является вопрос *создания единой ИБ системы лингвистического обеспечения* (второй, третий, четвертый уровни). В русле наметившихся *тенденций автоматизации функций информационного обеспечения*

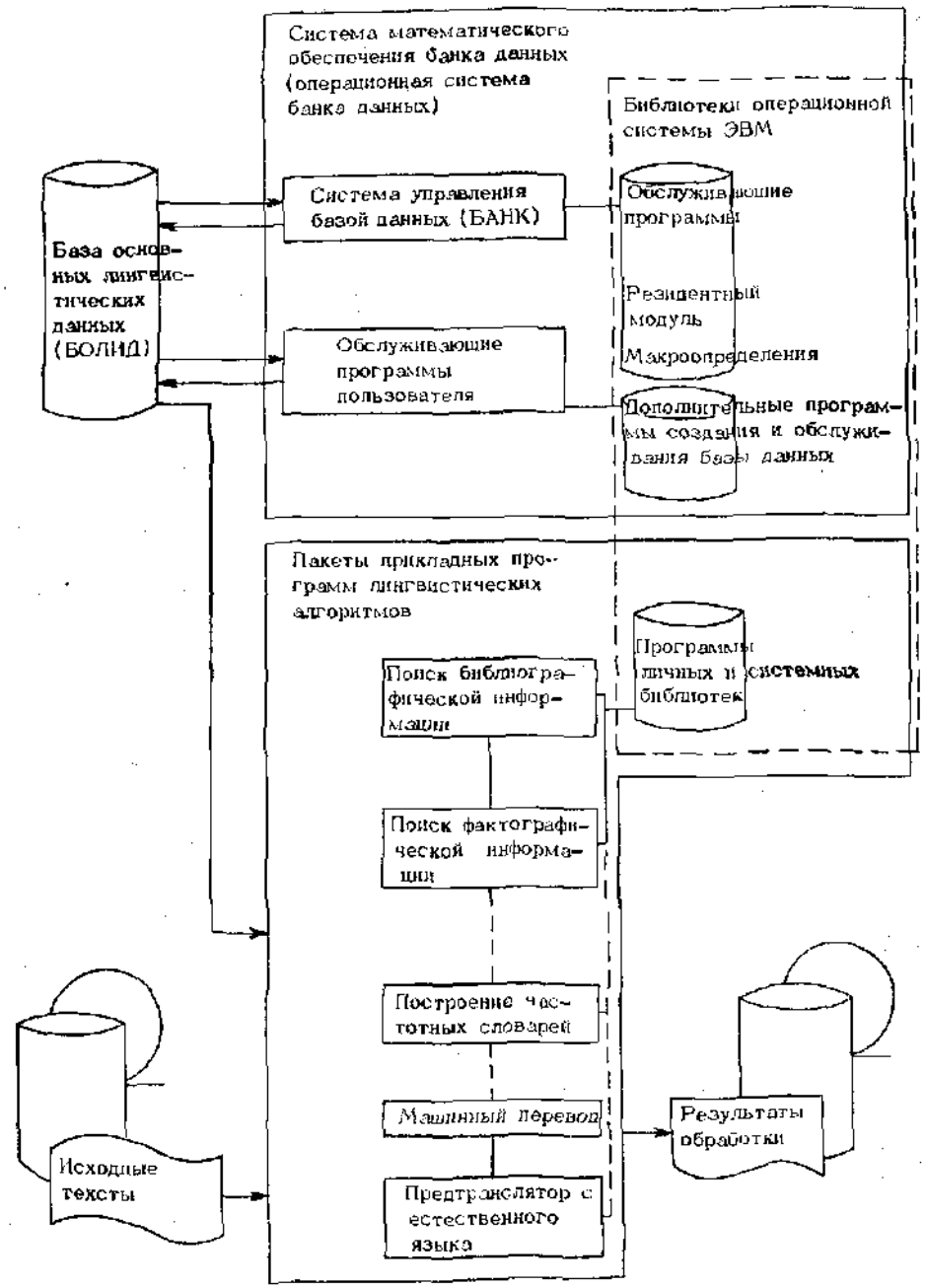


Рис. 1.1. Общая блок-схема системы лингвистического обеспечения.

проблема формирования информационной базы ЛО должна решаться с использованием банков данных. Банк лингвистических данных используется в общей системе ЛО согласно блок-схеме (см. рис. 1.1).

1.2. Информационная база лингвистического обеспечения (автоматические словари) и ее компонентная структура

Несмотря на то что машинная атрибуция, аннотирование и реферирование имеют целью извлечь из текста только общий смысловой инвариант, а машинный перевод, ориентированный на полную смысловую переработку текста, использует более сложные формы распознавания, охватывающие в идеале все единицы текста, все эти задачи можно решать, используя единую информационную базу. В общем случае информационная база класса лингвистиче-

	Грамматическая информация	Информация о терминологичности слова	Семантическая информация	Тезаурусная информация	Фразеологическая характеристика слова
ЛЕ	Содержит лексико-грамматические коды, определяющие синтактико-морфологические категории словоформы ЛЕ	Содержит коды подъязыков, в которых употребляется ЛЕ, и определяет терминологичность слова, а также принадлежность слова к эталонным спискам, т. е. дескрипторность	Содержит коды семантических классов ЛЕ внутри подъязыка	Содержит коды родовидовых и ассоциативных связей	Содержит ссылки на словарь оборотов через ядерные слова

Рис. 1.2. Структура словарной статьи АС.

ских задач представляет собой систему автоматических словарей (АС), состоящую из русского и иностранного словников, каждая из словарных статей (ССт) которых включает лексическую единицу (ЛЕ) с присущей ей тезаурусной, семантической, фразеологической, терминологической и грамматической информацией (см. рис. 1.2). Каждый АС условно разделяется на две части: общеупотребительную и терминологическую. При построении каждой части многоцелевого отраслевого АС общеупотребительная часть словника и словарь оборотов для любого языка будут оставаться неизменными, терминологическая же часть будет варьироваться в зависимости от тематики обрабатываемых текстов. Общеупотребительная лексика обнаруживается путем сопоставления по различным корреляционным критериям частот словоформ из общелитературных и отраслевых частотных словарей. Словоформы, показывающие достаточную корреляцию частот и равномерное распределение, Гаусса или Пуассона, а также общие значения для

всех исследованных подъязыков, включаются в словарь общеупотребительной лексики. Словоформы, присущие определенным подъязыкам, а также те словоформы, системы значения которых

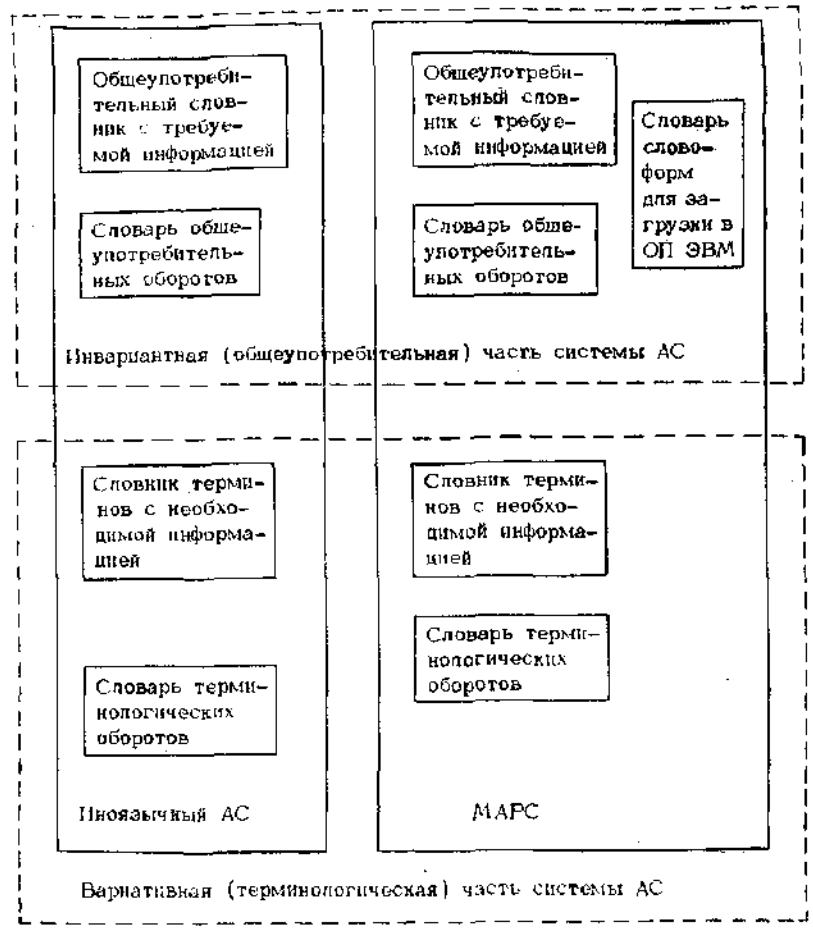


Рис. 1.3. Компонентная структура информационной базы лингвистического обеспечения.

не совпадают в различных подъязыках, группируются в терминологический словарь для данного подъязыка.

Система АС информационной базы ЛО обязательно содержит единый многоцелевой автоматический русский словарь (МАРС). Иноязычные АС включаются в ИБ в том случае, когда база должна обеспечивать выполнение лингвистических задач в двуязычной ситуации. Количество иноязычных словников в ИБ не ограничено (английский, немецкий, французский и т. д.). Все рассуждения далее проводятся относительно ИБ, содержащей русский

АС и один иноязычный АС (английский). Полученные результаты распространяются на произвольный состав ИБ. Компонентная структура ИБ ЛО представлена на рис. 1.3.

Так как вся информационная база из-за ее большого объема в основной памяти (ОП) ЭВМ размещена быть не может, для наиболее часто встречаемых служебных и некоторых других неизменяемых слов русского языка строится словарь словоформ (общим объемом 50 единиц), который вводится в память ЭВМ перед выполнением каждой лингвистической задачи. Порядок расположения ЛЕ в этом словаре — частотно-алфавитный. Остальная часть системы АС располагается на внешних носителях информации.

Информация включается в АС на основании частотных и алфавитно-частотных списков, которые составляются статистическим путем на основе массивов текста с помощью ЭВМ. При этом, если в информационную базу включен иноязычный АС, проверяется, имеются ли в русском словнике все требуемые для этого АС переводные эквиваленты. В случае необходимости русский словник дополняется. Так, для составления английского автоматического словаря по вычислительной технике были построены тексты длиной в 200 000 словоупотреблений, из которых 13 160 различных словоформ включены в английский словник.

Для составления русского АС по вычислительной технике из текстов объемом 250 000 словоупотреблений отобраны в словник 10 520 лексем. Как русский, так и иноязычный АС содержат словник и словарь оборотов.

Иноязычный словарь оборотов включает такие словосочетания, для которых в результате обращения к двуязычному словнику и грамматическим правилам анализа и синтеза невозможно найти правильный перевод на русский язык. Объем словаря оборотов составляет 1200 общеупотребительных оборотов и 3850 оборотов подъязыка вычислительной техники.

1.3. Банковское построение автоматических словарей

Автоматические словари информационной базы должны быть реализованы в виде системы наборов лингвистических данных и совокупности алгоритмических, программных и языковых средств, обеспечивающей порождение лексических единиц (морфем, словоформ, канонических форм, оборотов-сегментов), которые задаются соответствующим лингвистическим алгоритмом.

При реализации системы АС возникает дилемма: использовать для этой цели имеющиеся программные средства или разрабатывать новые. В этой связи необходимо отметить, что хотя существующие в настоящее время машинные АС системы МП и информационного поиска различных классов реализованы на базе ЭВМ второго поколения типа Минск-22, Минск-32, М-220, М-222, использовать их информационное программное (а заодно и жестко связанное с ними алгоритмическое) наследие сегодня не имеет смысла по следующим причинам.

1. Поиск и обработка данных в системах ЭВМ второго поколения, как известно, основаны на последовательном методе доступа, при котором для сокращения времени поиска требуются многократные сортировки.

2. Каждый массив данных жестко связан с программами, содержащими описание его структуры.

3. Потребность различных программ в обрабатываемых данных и их комбинациях весьма различна, что приводит к значительному увеличению числа массивов, с одной стороны, и неоднократному дублированию информации — с другой.

4. Требование обновления и пополнения системы АС приводит к трудоемкой операции реорганизации массивов.

Учитывая возможности и перспективы систем автоматизации информационного обеспечения ЭВМ третьего поколения, наиболее рационально реализовать всю информационную базу класса лингвистических задач в виде банка данных. Банк данных лишен перечисленных недостатков и характеризуется следующими особенностями:

1) БД расположен на устройствах прямого доступа и обеспечивает непосредственное обращение к информации базы;

2) создание, ведение и обновление данных автоматизировано и полностью отделено от программ;

3) в банке данных возможно хранение больших объемов данных и запоминание логических связей между ними, что позволяет исключить дублирование данных и упростить алгоритмы их обработки;

4) банк представляет собой открытую систему, что значительно облегчает пополнение базы новой информацией.

Основной удельный вес в банке лингвистических данных должен иметь МАРС.

Необходимость построения русского АС в виде многоцелевого универсального словаря объясняется следующими соображениями:

1. В подавляющем большинстве систем автоматической переработки текста, используемых в нашей стране, русский язык выступает не только в качестве выходного языка, но и в качестве метаязыка — эталона, относительно которого осуществляется описание входного языка (в том числе и самого русского). Поэтому при переработке текста на ЭВМ, осуществляемой в двуязычной ситуации, описание семантических пространств и их областей, а также некоторых операций по распознаванию смысла целесообразно осуществлять, насколько это возможно, средствами выходного (русского) языка независимо от типа входного языка по причине межъязыковой эквивалентности семантической и тезаурусной информации языка.

2. Как показывают информационные исследования текста (Пиотровский, 1975, с. 120—148), бóльшая часть синтаксической,

семантической и прагматической информации (до 80%), оказывается заложенной в лексике и фразеологии языка. В связи с этим целесообразно структурно и функционально заложить в универсальный русский словарь основную часть грамматической информации, необходимую для анализа и синтеза текста.

3. В настоящее время существует большое число автономных отраслевых АС русского языка для вероятностного детерминистского, тезаурусного распознавания смысла текста, свертывания текста и его машинного перевода, для реализации человеко-машинного диалога на естественном языке. Все эти словари целесообразно объединить и использовать для формирования единого многоцелевого автоматического русского словаря, о котором мы говорили выше.

Так же как разработка всего лингвистического обеспечения, создание системы АС связано с теоретическими препятствиями, возникающими при формализации лингвистического материала, а также с трудностями практической машинной реализации.

В настоящее время многие вопросы лингвистической теории АС еще окончательно не разработаны, нет четких семантических и тезаурусных классификаций. Тем не менее необходимость в создании единой ИБ класса ЛЗ диктуется нуждами развития информатики и управления: отсутствие в нашей стране практически действующих систем ЛО начинает тормозить разработку высокоэффективных АСУ и АИС, а отсутствие таких систем затрудняет в свою очередь дальнейшие лингвистические исследования в этом направлении.

Разработка единой информационной базы ЛО в структуре банка данных, наряду с указанными в 1.1 преимуществами БД, имеет свои нерешенные проблемы.

Создание банка данных является чрезвычайно трудоемкой задачей, результатом многолетней работы целого коллектива квалифицированных специалистов (Мартин, 1978, с. 13—20). Однако в данном случае разработки специального БД для реализации ИБ ЛО не требуется, поскольку сегодня в нашей стране и за рубежом для ряда информационных систем общего назначения создано математическое обеспечение, представленное в виде пакетов прикладных программ, и определены логические принципы их создания. Эти системы позволяют обеспечить реализацию некоторого базисного набора функций, гарантирующего основные требования пользователя к банкам данных. Так как системы, удовлетворяющие всем требованиям к банкам, в настоящее время еще не созданы, отсутствующие функции, равно как и дополнительные и специальные требования, должны обеспечиваться пользователем в зависимости от условий конкретных задач средствами программирования, предоставленными СУБД (стандартным математическим обеспечением банка), и ОС, в среде которой функционирует банк. В связи с этим для создания информационной базы системы лингвистического обеспечения должен быть определен конкретный банк данных,

разработанный в нашей стране, характеристики которого позволят реализовать логическую структуру базы данных, определяемую информационными моделями лингвистических задач. Так как базисный набор функций любого имеющегося банка не удовлетворяет требованиям к системе автоматизации информационного, а значит и лингвистического обеспечения, должны быть разработаны программные средства, расширяющие возможности банка в смысле повышения эффективности и улучшения качества выполнения разнообразных лингвистических задач АСУ и АИС.

1.4. Выводы

1. Проблема разработки лингвистического обеспечения есть проблема автоматизации одной из функций информационного обеспечения, а именно — автоматизации процессов переработки неформализованной информации. Эта актуальная проблема сталкивается с недостаточностью разработки лингвистических теорий АС, с одной стороны, и трудностями машинной реализации с целью автоматизации функций информационного обеспечения — с другой.

2. Разработка открытой системы лингвистического обеспечения позволит по мере достижений лингвистической теории обновлять, пополнять, модифицировать систему ЛО. Это даст возможность пользователю различных вычислительных систем уже сегодня работать в условиях «удобной вычислительной обстановки» на базе достигнутого и определит новый импульс для теоретических исследований в области прикладной и инженерной лингвистики.

3. При программной реализации лингвистические алгоритмы должны быть представлены в виде пакетов прикладных программ, имеющих единую информационную базу.

4. Создание, ведение, поддержание информационной базы в актуальном состоянии должно осуществляться средствами банков данных, как одним из самых мощных орудий автоматизации функций информационного обеспечения.

5. Основой информационной базы ЛО являются автоматические словари, наибольший удельный вес среди которых должен иметь универсальный многоцелевой автоматический русский словарь. Его универсальность определяется тем, что при всех видах автоматической переработки неформализованных текстов, в том числе и иностранных, используется единая лексическая и грамматическая информация, которая и закладывается в МАРС. В настоящее время такого словаря не существует, он должен создаваться на основе имеющихся отдельных словарей автоматизированных систем научно-технической информации, различных толковых, энциклопедических, терминологических словарей и содержать грамматическую, семантическую, тезаурусную информацию. Как компонент ЛО, словарь должен представлять собой открытую подсистему.

6. При машинной реализации системы АС, согласно принятой концепции, должны использоваться существующие в нашей стране банки данных. Недостающие пользователю функции банка необходимо реализовать в виде комплекса обслуживающих программ.

§ 2. Проблема структурирования базы основных лингвистических данных

Система автоматических словарей, являющаяся ядром лингвистического обеспечения, должна быть универсальной, т. е. обеспечивать с информационной точки зрения решение рассматриваемого в 1.1 класса ЛЗ. В связи с этим система АС должна описывать все многообразие функционирования лексических единиц (в частности, каждая ЛЕ русского и иноязычного АС должна характеризоваться сложной совокупностью присущих ей структурных признаков). Это означает, что в системе АС, которую далее будем именовать базой основных лингвистических данных (БОЛИД), должна поддерживаться в актуальном состоянии такая сложная информационная модель понятий и объектов внешнего мира и их взаимоотношений, как естественный язык.

Как показано выше (см. 1.3), такая модель наиболее оптимально реализуется в структуре банка данных, который обеспечивает хранение массивов большого объема и фиксацию сложных связей между отдельными компонентами этих массивов.

При реализации банка лингвистических данных возникает ряд вопросов, решение которых в конечном счете определяет систему управления данными, структуру данных, физическую организацию данных в БОЛИД. Рассмотрим эти вопросы более подробно.

2.1. Основные понятия

Прежде чем приступить к постановке задачи, введем некоторые понятия и определения, важные для последующего изложения (ср.: Информационные системы общего назначения, 1975).

1. Элемент данных (ЭД) — элементарная, логически неделимая единица данных. Из ЭД составляются структуры всех остальных типов.

2. Агрегат данных (АД) — набор элементов данных, логически представляющий собой совокупность свойств некоторого реального объекта.

3. Отношение — логическая связь между двумя объектами или их отдельными свойствами, представленными АД (ЭД). Выделены четыре типа отношений между АД (ЭД): однооднозначные, одномногозначные, многооднозначные и многомногозначные (см. рис. 2.1).

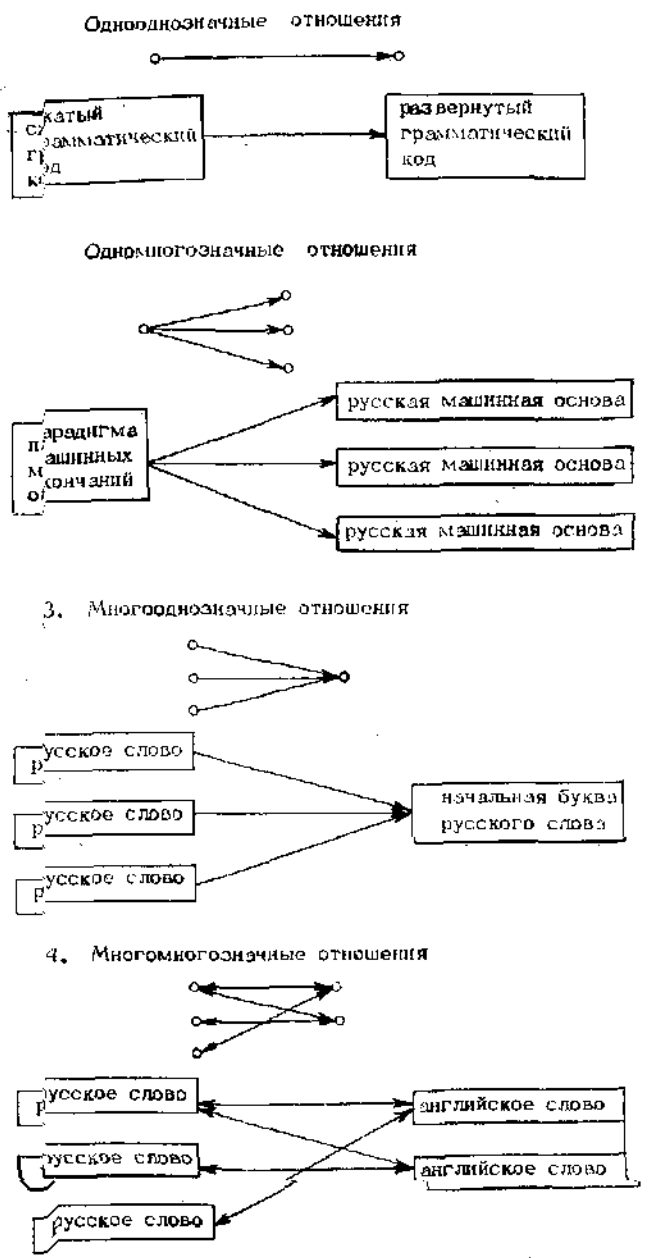


Рис.2.1. Примеры отношений в системе АС.

4. Ключ поиска (КП) — элемент данных или агрегат данных, определяющий перечень исходных свойств объекта и необходимый для поиска объекта в ИБ.

5. Поисковая процедура (ППр) — последовательность обращений к ИБ, необходимых для размещения, поиска или корректировки какого-либо элемента данных по ключу поиска. Каждая ППр представляет собой иерархический граф с одной корневой вершиной (см. рис. 2.2).

6. Лингвистическая задача — набор процедур поиска и обработки данных, в процессе выполнения которых производится один из видов автоматической переработки неформализованных текстов согласно 1.1.

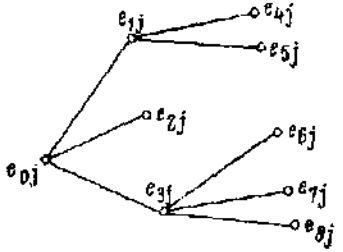


Рис.2.2. Граф j-той поисковой процедуры.

e_{0j} — корневая вершина; $\{e_{1j}, e_{2j} \ldots, e_{8j}\}$ — элементы данных, которые могут играть роль искомых; $\langle e_{0j}\rangle$, $\langle e_{0j}, e_{1j}\rangle$, $\langle e_{0j}, e_{2j}\rangle$ — элементы и агрегаты данных, которые могут служить ключами поиска.

Рис. 2.3. Процедурная структура лингвистической задачи.

Данное определение предполагает, что лингвистическая задача, как и любая другая задача, которая подвергается машинной реализации, может быть представлена в виде последовательности отдельных процедур, которые в общем случае делятся на два типа: процедуры поиска и процедуры обработки (см. блок-схему на рис. 2.3). Для каждой ЛЗ рассматриваемого класса определяется множество ППр $\{A_j^s\}$, $j=\overline{1, J}$ таким образом, что для всего класса ЛЗ можно считать заданным множество поисковых процедур $\{A_j\}$, $j=\overline{1, J}$.

7. Запись (группа) — наименьшая структурная (логическая) единица информационной базы, содержащая один или несколько АД (ЭД). Группа есть древовидный граф с одной корневой вершиной, соответствующей ключевому элементу группы. Ключевой элемент группы — корневая вершина иерархической структуры, которая соответствует некоторому объекту, а подчиненные вершины — его свойствам и другим объектам, связанным с основным объектом отношениями.

8. Цепь (групповое отношение) — логическая связь между двумя АД (ЭД), принадлежащими равным группам.

9. Поле данных — совокупность байтов (символов), физически отображающих элементы данных.

10. Страница (физическая запись, блок) — единица обмена информацией между ИБ и основной памятью ЭВМ.

11. Файл — набор физических записей, организованный стандартным для системы образом на магнитных дисках.

2.2. *Содержательная постановка задачи*

Система автоматической переработки неформализованных текстов, как и всякая автоматизированная система, использует некоторое пространство данных, содержащее множество элементов (ЭД и АД) и отношений между этими элементами. Будем называть такое пространство элементным пространством (ЭПр).

Элементное пространство определяется областями, используемыми в различных поисковых процедурах, входящих в набор лингвистических задач. Каждая поисковая процедура может быть представлена некоторым графом (см. рис. 2.2), вершинами которого являются элементы, а дугами — отношения между ними. Каждой вершине графа можно поставить в соответствие множество свойств данного элемента, а каждое дуге — множество свойств данного отношения.

Тогда всё ЭПр может быть представлено графом, объединяющим графы отдельных областей данных. Так как между двумя элементами в ЭПр для различных ППр может существовать несколько типов отношений, в общем случае элементное пространство представляется мультиграфом.

Всякое отношение между двумя элементами ИБ реализуется тремя возможными способами:

1) физической группировкой этих элементов в памяти;

2) указанием адресных ссылок для элементов, расположенных в памяти автономно;

3) алгоритмическим или программным путем (также при автономном расположении элементов в памяти). При структурировании информационной базы данных интерес представляют первые два типа реализации, так как именно они предполагают отображение отношения в логическую, а следовательно, и физическую структуры данных.

Будем называть логической структурой данных ИБ совокупность логических единиц — элементов данных, агрегатов данных, групп, групповых отношений, составляющих ИБ и рассматриваемых безотносительно к средствам и методам их хранения в памяти ЭВМ. Под физической структурой ИБ понимается совокупность физических единиц хранения данных — полей, физических записей, файлов — с присущими им параметрами физической организа-

ции. Между логической структурой данных и их физической организацией существует определенная функциональная связь.

При структурировании ИБ тот или иной способ реализации отношений выбирается с учетом двух требований: минимизации объема памяти, занимаемой базой на внешних устройствах, и минимизации времени поиска информации в базе. Каждый из этих критериев зависит от совокупности параметров, определяющих логическую и физическую структуру базы; при этом оба требования являются альтернативными (см. рис. 2.4). Отсюда следует, что одновременная минимизация двух критериев не может быть достигнута. Поэтому вводится понятие суммарной минимизации ресурсов.

Существо такой минимизации состоит в том, чтобы для удовлетворения обоих альтернативных требований найти минимальное значение некоторой функции $P(V, T)$ (см. рис. 2.4), такой, что

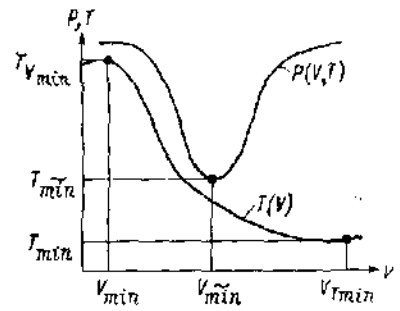


Рис. 2.4. Графическое определение точки суммарной минимизации ресурсов.

$$P(V, T) = C_1 \cdot C_V \cdot V + C_2 \cdot C_T \cdot T,$$

где C_1 и C_2 — нормирующие коэффициенты, необходимые для приведения параметров к безразмерной форме, они могут быть заданы как величины, обратные максимально допустимым значениям объема V и времени поиска T: $C_1 = 1/V_{max}$, $C_2 = 1/T_{max}$,

C_V и C_T — весовые коэффициенты, величина которых определяется специалистом — постановщиком задачи при условии, что $C_V + C_T = 1$.

Таким образом, в общем виде задача структурирования БОЛИД сводится к нахождению точки минимума значения функции $P(V, T)$, которую назовем точкой суммарной минимизации ресурсов.

Эта точка $(V_{\tilde{min}}, T_{\tilde{min}})$ характеризуется некоторой определенной совокупностью логических и физических параметров базы, которые и определяет оптимальное сочетание объема $V_{\tilde{min}}$ и времени поиска информации в базе $T_{\tilde{min}}$ (см. рис. 2.4).

2.3. Математическая постановка задачи структурирования ИБ

Известны:

I. Множество элементов (ЭД, АД) информационной базы

$$E = \{e_i, P_i\}, \quad i = \overline{1, I},$$

где e_i — имя элемента,

P_i — совокупность параметров элемента

$$P_i = [p_i, h_i],$$

где p_i — длина поля элемента i-того типа,

h_i — количество значений элемента i-того типа.

Тогда каждая пара $\langle e_i, p_i \rangle$ может быть представлена следующим образом:

$$\langle e_i, p_i \rangle = e_{i_1}, p_i \cup e_{i_2}, p_i \cup \ldots \cup e_{i_{h_i}}, p_i,$$

где e_{i_k} — значение i-того элемента.

II. Множество поисковых процедур класса лингвистических задач

$$A = \{a_j, P_j\}, \quad j = \overline{1, J},$$

где a_j — имя ППр,

P_j — совокупность параметров ППр.

Совокупность параметров P_j определяется как

$$P_j = \{e_j^a, f_j, \varphi_j\}, \quad j = \overline{1, J},$$

где e_j^a — подмножество множества элементов E информационной базы, используемых в j-той ППр.

Для определения функций f_j и φ_j введем следующие понятия.

Пусть имеется:

1) ρ — множество соответствий между парами множеств элементов e_i и e_l:

$$e_i, e_l \in e_j^a,$$

$$\mathrm{P} = \{\rho_{il}\}.$$

2) W — множество типов соответствий:

$$W = \{w_1, w_2, w_3, w_4\}.$$

Причем, $w_1 \subset W$ состоит из всех однозначных отображений e_i на e_l.

$w_2 \subset W$ состоит из всех типов отображений e_i на e_l, не являющихся взаимнооднозначными.

$w_3 \subset W$ состоит из тех и только тех соответствий e_i и e_l, при которых каждому значению элемента e_i соответствует несколько значений элемента e_l и для каждого значения $e_{l_k} \in e_l$ найдется единственное значение элемента $e_{i_\tau} \in e_i$, такое, что e_{i_τ} соответствует e_{l_k}, причем для некоторого $e_{i_\tau}^{(0)} \in e_i$ существует несколько e_{l_z}, ему соответствующих.

$w_4 \subset W$ состоит из тех и только тех соответствий e_i и e_l, при которых каждому значению элемента из e_i соответствует одно или несколько значений элементов из e_l; для каждого значения элемента $e_{l_t} \in e_l$ найдется одно или несколько значений элемента

$e_{i_\tau} \in e_i$, таких, что e_{i_τ} соответствует e_{l_ξ}, причем для некоторого e_{i_τ} существует несколько e_{l_ξ}, ему соответствующих, и для некоторого $e_{l_\xi}^{(0)} \in e_l$ существует несколько значений $e_{i_\tau} \in e_i$, таких, что e_{i_τ} соответствует $e_{l_\xi}^{(0)}$.

Тогда функции f_j и φ_j определяются на декартовом произведении элементов из e_j следующим образом:

$$f_j : [1, 2, 3, \ldots, i] \times [1, 2, 3, \ldots, \tilde{I}] \to P$$

$$\text{или } f_j(i, l) = \rho_{il},$$

$$\varphi_j : [1, 2, 3, \ldots, \tilde{I}] \times [1, 2, 3, \ldots, \tilde{I}] \to W$$

$$\text{или } \varphi_j(i, l) = \begin{cases} \alpha, & \text{если } \rho_{il} \in w_1 \\ \beta, & \text{если } \rho_{il} \in w_2 \\ \gamma, & \text{если } \rho_{il} \in w_3 \\ \delta, & \text{если } \rho_{il} \in w_4 \\ \varnothing, & \text{если } \rho_{il} \notin W. \end{cases}$$

Коды α, β, γ, δ, $\varnothing$ определяют следующие типы соответствий между элементами e_i и e_l j-той поисковой процедуры (см. рис. 2.1):

α — одноднозначное соответствие,
β — одномногозначное соответствие,
γ — многоднозначное соответствие,
δ — многомногозначное соответствие,
$\varnothing$ — *отсутствие отношений (соответствия).*

Требуется определить структуру данных и структуру хранения информационной базы основных лингвистических данных.

I. Множество групп (записей)

$$R = \{r_\eta,\ P_\eta\},\quad \eta = \overline{1,\ H},$$

где r_η — тип группы,
P_η — совокупность параметров группы.

$$P_\eta = [\{e_i\}_\eta,\ \{\hat{e}_i\}_\eta,\ p_\eta,\ \pi_\eta,\ h_\eta],$$

где $\{e_i\}_\eta$ — множество ЭД (АД), входящих в группу η-типа,
$\{\hat{e}_i\}_\eta$ — множество служебных ЭД, входящих в группу η-типа,
p_η — длина группы данного η-типа,
π_η — способ поиска (размещения) записи типа η в информационной базе

$$\pi_\eta \in \Pi,$$

где Π — множество способов поиска (размещения) записей в информационной базе, определяемое конкретной СУБД. Этот параметр определяется на этапе построения физической структуры базы данных,
h_η — количество экземпляров группы η-типа.

II. Множество связующих типов (записей)

$$R^c = \{r_\nu^c,\ P_\nu^c\},\quad \nu = \overline{1,\ N},$$

где r_ν^c — имя связующего типа записи,
P_ν^c — совокупность параметров записи связи типа ν.

$$P_\nu^c = [\{e_i\}_\nu,\ \{\hat{e}_i\}_\nu,\ p_\nu,\ \pi_\nu,\ h_\nu],$$

где $\{e_i\}_\nu$ — множество информационных элементов базы, включенных в запись связи типа ν (это множество может быть пусто),
$\{\hat{e}_i\}_\nu$ — множество служебных полей записи связи ν-типа,
p_ν — длина записи связи ν-типа,
π_ν — способ поиска (размещения) записи типа ν в информационной базе, $\pi_\nu \in \Pi$,
h_ν — количество экземпляров записей связи типа ν.

III. Множество логических связей (цепей, групповых отношений)

$$L = \{l_\xi,\ P_\xi\},\quad \xi = \overline{1,\ \Xi},$$

где l_ξ — имя цепи,
P_ξ — совокупность параметров цепи.

$$P_\xi = [w_\xi,\ \langle e_{i_\xi},\ e_{l_\xi}\rangle],$$

где w_ξ — тип отношения, $w_\xi \in W$ (этот параметр является исходным),
$\langle e_{i_\xi},\ e_{l_\xi}\rangle$ — пара элементов информационной базы, связанных цепью l_ξ.

IV. Множество полей данных

$$Q = \{q_\mu,\ P_\mu\},\quad \mu = \overline{1,\ M},$$

где q_μ — имя поля μ-типа,
P_μ — совокупность параметров поля μ-типа.

$$P_\mu = [\{e_i\}_\mu,\ f_\mu^q],$$

где $\{e_i\}_\mu$ — множество элементов ИБ, входящих в состав поля μ-типа,
f_μ^q — формат представления данных в поле μ-типа.

V. Множество файлов информационной базы

$$\Phi = \{\varphi_\chi,\ P_\chi\},\quad \chi = \overline{1,\ X},$$

где φ_χ — имя файла,
P_χ — совокупность параметров, характеризующих файл типа χ.

$$P_\chi = [\{r_\eta\}_\chi,\ \{r_\nu\}_\chi,\ p_\chi],$$

где $(r_\eta)_\chi$ — множество связующих типов записей, входящих в файл типа χ,

$(r_\nu)_\chi$ — множество типов записей, входящих в файл типа χ,

p_χ — размер страницы (блока, физической записи) файла типа χ.

Управляемые параметры R, R^e, L, Q, Ф должны быть определены таким образом, чтобы объем базы БОЛИД V и время поиска информации в ней T, которые являются основными ресурсами информационной базы, были минимальны. Иными словами, требуется минимизировать функционалы

$$V = F_1\{(e_i)_\eta,\ p_\eta,\ h_\eta,\ (\hat{e}_i)_\eta,\ (e_i)_\nu,\ p_\nu,\ h_\nu,\ (\hat{e}_i)_\nu\}$$

и

$$T = F_2\{(e_i)_\eta,\ p_\eta,\ h_\eta,\ (\hat{e}_i)_\eta,\ (e_i)_\nu,\ p_\nu,\ h_\nu,\ (\hat{e}_i)_\nu,\ (e_i)_\mu,\ f^\eta_\mu,\ (r_\eta)_\chi,\ p_\chi,\ \pi_\eta,\ \pi_\nu\}.$$

Так как, согласно 2.2, одновременная минимизация F_1 и F_2 невозможна, рассмотрим функцию $P(V, T)$, которая может быть также представлена как некоторый функционал

$$P(V, T) = F\{(e_i)_\eta,\ p_\eta,\ h_\eta,\ (\hat{e}_i)_\eta,\ (e_i)_\nu,\ p_\nu,\ h_\nu,\ (\hat{e}_i)_\nu,\ (e_i)_\mu,\ f^\eta_\mu,\ (r_\eta)_\chi,\ p_\chi,\ \pi_\eta,\ \pi_\nu\},$$

где $\eta = \overline{1, \text{N}}$, $\nu = \overline{1, \text{N}}$, $\xi = \overline{1, \text{E}}$, $\mu = \overline{1, \text{M}}$, $\chi = \overline{1, \text{X}}$.

Минимизировать этот функционал в общем виде, т. е. решить комбинаторную задачу огромной размерности и сложного нелинейного вида, не представляется возможным ни точным, ни приближенным методом; единственным приемлемым путем решения данной задачи является эвристический способ поиска некоторого рационального решения, удовлетворяющего требованиям как по объему ИБ, так и по времени работы с ней.

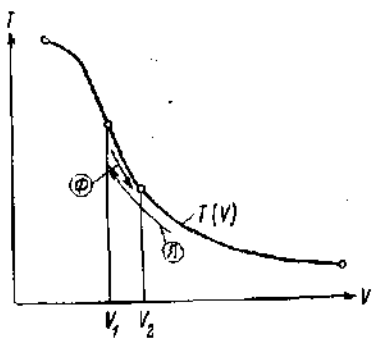


Рис. 2.5. Последовательное приближение к точке суммарной минимизации ресурсов при структурировании БОЛИД.

Л — логическое структурирование базы БОЛИД, Ф — физическое структурирование базы БОЛИД.

Методика нахождения рационального технического решения в данном случае будет определяться тремя последовательными этапами.

Первый этап формирования ИБ АС — это этап определения элементной структуры АС, который реализуется по критерию минимизации объема системы автоматических словарей при определении лексической единицы и состава словарной статьи АСл.

Второй этап соответствует синтезу структуры данных ИБ. На этом этапе вводится понятие оценочной функции. Использование этого критерия при структурировании данных позволяет минимизировать объем необходимой и служебной информации в базе.

На третьем этапе — при реализации структур хранения, т. е. физической организации данных в ИБ — учитывается ряд факторов, уменьшающих время поиска информации в базе. Естественно, что при этом наблюдается некоторое увеличение объема базы, допустимое, если объем V не превышает объема одновременного доступа к информации в БОЛИД.

Графически этапы логического и физического структурирования данных относительно времени поиска информации в базе показаны на рис. 2.5.

Такой подход к решению поставленной задачи объясняется тем, что объем базы в основном зависит от ее логической структуры, тогда как время поиска информации в базе в значительной мере зависит и от ее физической организации.

2.4. *Выводы*

1. При постановке задачи структурирования базы основных лингвистических данных в качестве исходных параметров выступают множество элементов структуры автоматических словарей и множество поисковых процедур рассматриваемого класса ЛЗ.

2. Структурирование БОЛИД распадается на три этапа: определение элементной структуры АС, логическое структурирование и физическое структурирование базы лингвистических данных.

3. На этапе логического структурирования управляемыми параметрами являются группы и групповые отношения в базе, на этапе физического структурирования — поля, физические записи и файлы.

4. Управляемые параметры должны быть определены таким образом, чтобы объем ИБ и время поиска информации в ней были минимальны. Так как в общем виде решить задачу суммарной минимизации ресурсов ИБ не представляется возможным, предлагается эвристический способ поиска некоторого рационального решения, удовлетворяющего как по объему базы, так и по времени работы с ней.

5. Рациональное решение поставленной задачи включает последовательную минимизацию объема базы на этапе определения элементной структуры АС и на этапе формирования структур данных, минимизацию времени поиска информации в базе на этапе физической организации данных.

§ 3. Разработка элементной структуры автоматических словарей

Согласно компонентной структуре информационной базы (рис. 1.3), автоматический словарь, независимо от того, является он русским или иноязычным, состоит из общеупотребительной и терминологической частей, каждая из которых в свою очередь делится на два компонента: словник АС, т. е. список лингвистических единиц словаря (словоформ или основ) со всей относящейся к нему информацией (см. рис. 1.2), и словарь оборотов.

Очевидно, что элементную структуру словаря, т. е. множество $E=\{e_i\}$, однозначно определяют следующие показатели: тип лексической единицы, состав словарной статьи, методы организации лексико-грамматической, семантической, терминологической, тезаурусной, фразеологической характеристик слова. Все перечисленные параметры структуры АС, согласно постановке задачи, должны определяться с учетом требования минимизации объема базы.

3.1. Оценка объема лексики в автоматических словарях

Автоматические словники ИБ создаются путем объединения более мелких словарных массивов в единые унифицированные макрословари, доступные для разнообразных приложений.

Формирование словарей осуществляется путем объединения имеющихся словников и списков устойчивых словосочетаний из автоматических двуязычных отраслевых словарей со словниками тезаурусов, толковых, энциклопедических, обратных и других словарей. Прежде чем говорить о структуре словарной статьи автоматических словарей, произведем оценку объема лексики АС. Для этого воспользуемся лексикографическими данными по английскому языку, исходя из того, что в англо-американской лексикографии получено наиболее полное описание словарного состава языка.

Третье издание словаря Уэбстера (Webster, 1961) содержит около 450 тыс. словарных статей. Современные американские тезаурусы по различным областям знаний включают от 3 до 14 тыс. дескрипторов и ключевых слов каждый, причем дескрипторы и ключевые слова фильтруются из списков, содержащих многие десятки тысяч простых и сложных терминов, часть из которых отсутствует в словаре Уэбстера. Таким образом, можно предположить, что весь корпус английской лексики, включая терминологические слова и словосочетания, будет приближаться к 1 млн. лексических единиц.

Обратимся к данным русских словарей. «Обратный словарь русского языка» (ОСл, 1974), обобщивший словники четырех словарей русского литературного языка, включает около 125 тыс. слов, в число которых вошло сравнительно небольшое количество специальных терминов. «Фразеологический словарь русского языка» и «Русско-английский словарь идиом» содержат около 10 тыс. общеупотребительных фразеологизмов (Беляева, Лукьянова, Пиотровский, 1976, с. 28—46). Каждая из терминологий отдельных отраслей знаний содержит многие тысячи лексических единиц. Так, например, «Словарь дескрипторов по химии и химической промышленности» содержит 20 тыс. дескрипторов и 8 тыс. синонимов, извлеченных из корпуса химических и нефтехимических терминов общим объемом в 180 тыс. простых и сложных терминов, многие из которых не засвидетельствованы в словарях русского литературного языка (Беляева, Лукьянова, Пиотровский, 1976, с. 28—46). Аналогичную картину показывают терминологические словари-справочники. Результаты предварительного анализа общего объема различных слов, зафиксированные в таких изданиях, дают число, превышающее 800 тыс. ЛЕ. В итоге при оценке словарного состава русского языка можно также прийти к объему в 1 млн. лексических единиц.

Английский язык является флективно-изолирующим языком, поэтому количество ССт в АСл английского АС близко к лексическому объему словаря. Русский же язык — язык флективный. Согласно 1.3, основная тезаурусная, грамматическая и семантическая информация относится к русскому АС независимо от того, какая ситуация используется при решении задачи: одно- или двуязычная. Поэтому, считая, что средняя длина английской словоформы равна 6 символам, можно определить объем английского словаря приблизительно в 15 млн. символов.

Если же представить, что словник МАРС будет строиться в виде обычного словаря словоформ, а русская ЛЕ имеет в среднем 10 форм слова, и длина каждой словарной статьи в среднем равняется 35 символам, общий объем словаря будет равен 350 млн. символов.

Так как наши словари являются словарями автоматическими, т. е. они должны быть расположены в памяти ЭВМ (основной или внешней), сразу становится ясной необходимость их компрессии при определении элементной структуры словарей, ибо внешняя память ЭВМ третьего поколения вычислительной системы не сможет вместить словари указанного объема на несменных носителях информации. А ведь пока речь шла только об объеме лексики автоматических словарей без учета объемов памяти, занимаемых грамматической, терминологической, фразеологической, тезаурусно-семантической информацией, без учета объема словаря оборотов.

Естественно, что в связи с тем, что объем русского АС даже по предварительным расчетам более чем в десять раз превосходит объем английского АС, основное внимание далее уделяется компрессии элементов русского автоматического словаря.

3.2. Выбор лексической единицы АС по критерию минимального объема информационной базы

Словники АС можно представить себе как наборы гнезд, в каждом из которых содержится последовательность форм одного и того же слова. Представление гнезда в словаре может быть задано тремя способами:

σ — изолирующим, когда в гнезде фиксируется вся парадигма одной ЛЕ в одном виде (см. рис. 3.1, *а*);

α — флективным, когда одной основе соответствует заданный в явном виде (в гнезде) список окончаний (см. рис. 3.1, *б*);

ι — агглютинативным, когда основе соответствует номер (адрес) парадигмы окончаний (см. рис. 3.1, в и 3.1, г).

Поставим в соответствие трем типам словарей два типа машинной морфологии: изолирующему типу — изолирующую машинную морфологию, агглютинативному и флективному типам словарей — агглютинативную машинную морфологию.

Рассмотрим, как решается проблема машинной морфологии в словарях всех трех типов, ибо основным доводом, который приводится в пользу словаря словоформ, является простота изолирующей машинной морфологии при выполнении различных лингвистических алгоритмов.

I. В словаре σ-типа каждой словоформе приписывается свой сжатый код (СК) лексико-грамматической информации. По совпадению этого кода и СК лексико-грамматической информации, определенного в результате анализа ЛЕ иноязычного текста, происходит отождествление словоформ в двуязычной ситуации. В одноязычной ситуации сличение слова текста со словоформами словаря осуществляется до их полного совпадения.

Очевидно, что такой тип машинной морфологии является предельно простым.

II. Машинная морфология для словаря ι-типа имеет следующий вид.

1. Каждой парадигме типовых окончаний приписываются номер в массиве окончаний и инициальный СК, характеризующий каноническую форму слова.

2. Согласно стандартной схеме лексико-грамматического кодирования, каждое поле парадигмы будет иметь СК, на единицу больший предыдущего (но только внутри парадигмы). Так, для существительного с номером парадигмы окончаний n и инициальным сжатым кодом k_j все двенадцать полей парадигмы имеют СК:

$$k_j,\ k_j+1,\ k_j+2,\ \ldots,\ k_j+12.$$

Выбор поля, т. е. выбор соответствующего окончания к машинной основе (МО) в двуязычной ситуации происходит по СК лексико-грамматической информации k'_j, определенному в результате анализа слова иноязычного текста. Номер n типовой парадигмы

а

Словообразовательное гнездо для словаря изолирующего типа	
Русский АС	Английский АС
система системы системе систему системой системе системы систем системам системы системами системах	system systems

б

Словообразовательное гнездо для словаря флективного типа	
Русский АС	Английский АС
систем а ы е у ой е ы — ам ы ами ах	system — *s*

в

Словообразовательное гнездо для словаря агглютинирующего типа	
Русский АС	Английский АС
систем	system

г

Наборы машинных окончаний к словообразовательному гнезду словаря агглютинативного типа	
Русский АС	Английский АС
а ы е у ой е ы — ам ы ами ах	— s

Рис. 3.1. Структура словообразовательного гнезда для трех типов автоматических словарей.

f_n, приписанный МО АС, определяет местоположение типовой парадигмы f_n в массиве окончаний. Номер поля в парадигме j определяется как

$$j = k'_j - k_j.$$

Конкретное машинное окончание, находящееся в поле с номером j, присоединяется к имеющейся машинной основе.

В одноязычной ситуации происходит обратный процесс: сличение слова текста и машинной основы из ССт. При частичном совпадении цепочки букв от начала слова с какой-либо МО словаря в парадигме окончаний, характеризующей данную основу, отыскивается поле, совпадающее по значению с оставшейся цепочкой букв слова. В случае полного совпадения словоформа текста считается отождествленной. Порождение канонической формы происходит путем присоединения к машинной основе словаря содержимого первого поля (поля с номером $j=1$) парадигмы, соответствующей МО.

III. Словарь α-типа имеет аналогичный алгоритм машинной морфологии, который несколько упрощается за счет того, что парадигма окончаний, соответствующая данной МО, находится непосредственно в каждой ССт словаря, и обращаться к массиву окончаний по номеру n для отыскания требуемой парадигмы не приходится.

Критерием выбора типа русского АСл является его минимальный объем. В связи с этим произведем сравнение словарей всех трех типов по объему.

Объем словника σ-типа определен в 2.1. Объемы словников α-типа и ι-типа определяются как соответственно в 3—5 и 5—7 раз меньше словника σ-типа. Проанализируем это положение.

Определим словари α-типа и ι-типа как словари основ. При этом имеется в виду не традиционная, а машинная основа (МО), т. е. цепочка букв от начала слова, общая для всех словоформ данной парадигмы (ср.: Беляева, Лукьянова, Пиотровский, 1976, с. 28—44). Цепочка букв, оставшаяся после выделения МО, рассматривается как машинное окончание.

Тогда в словарях α-типа и ι-типа вместо набора ССт в гнезде будет присутствовать одна ССт, содержащая МО этого гнезда (а не все гнездо словоизменений одного слова).

Чтобы унифицировать описание лексемы в словарях α-типа и ι-типа, введем два дополнительных грамматических понятия: нулевая парадигма и нулевая основа. Нулевая парадигма приписывается несклоняемым существительным типа *реле*, *ЭВМ*, а также наречиям, предлогам, союзам и частицам, не вошедшим в список слов, помещаемый в основную память. В этих случаях МО равна самому слову. Понятие нулевой основы используется при представлении слов с супплетивной парадигмой (ср. *он*, *его*, *ему*, *им* . . .; *идти*, *иду*, . . ., *шел* . . .). В этих случаях считается, что ССт содержит нулевую МО, к которой присоединяются машинные окончания, причем в качестве последних выступают супплетивные формы типа *он*, *его* . . ., *идти*, *шел* и т. д. Порождение нужной словоформы осуществляется стандартной операцией аппликирования машинного окончания к МО. В словаре α-типа явно указываются все машинные окончания в гнезде одной машинной основы, являющейся их владельцем. Такое сжатие, как известно, носит название сжатия парадигмы по методу Купера.

Словарь ι-типа строится из словаря α-типа путем вынесения парадигм машинных окончаний из ССт словаря в специальный массив машинных окончаний.

На основании вышесказанного ясно, что английский АСл должен быть представлен в ИБ как словарь изолирующего типа,

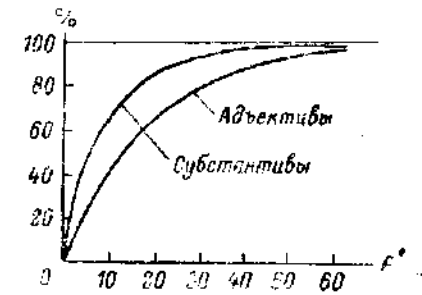


Рис. 3.2. Процентное распределение типовых парадигм по словарю в порядке убывания частоты их встречаемости.

ибо среднее число словоформ в гнезде ЛЕ английского языка не превышает 3, а машинная морфология для словарей σ-типа является предельно простой.

Для русского АС очевидным является лишь тот факт, что в каждом гнезде одной ЛЕ вместо набора словарных статей будет присутствовать одна ССт, содержащая машинную основу этого гнезда. Тогда общее количество ССт в словаре уменьшается приблизительно в 10 раз. В словаре ι-типа уменьшить объем можно также и за счет компрессии массива машинных окончаний. Этот массив компрессируется за счет наличия омографии (т. е. совпадения) парадигм машинных окончаний различных слов русского языка. Дело в том, что наборы машинных окончаний (далее будем именовать их просто окончаниями), соответствующие парадигмам различных слов русского языка, часто совпадают. Поэтому целесообразно выделить на всей совокупности наборов окончаний множество наборов $F=\{f\}$ и определить его как множество типовых парадигм, характеризующих совокупности различных лексем. Типовые парадигмы распределены по словарю неравномерно: анализ 5894 различных существительных подъязыка вычислительной техники показал, что 10 наиболее частотных парадигм охватывает около 70% всех существительных, первые 20 парадигм покрывают 90% существительных. Подобная же картина наблюдается и для других частей речи (см. рис. 3.2).

Определим объем словаря каждого типа, введя следующие обозначения:

V — объем словаря (в символах);

b_i — длина i-той МО (в символах);

p_k — длина k-того окончания (в символах);
$\hat{I}$ — общее число лексем в словаре;
l — среднее число словоформ одной лексемы;
N — число грамматических классов;
M — число типовых парадигм окончаний данного грамматического класса слов.

Тогда

$$V_\sigma = \sum_{i=1}^{\hat{I}} \left(l \cdot b_i + \sum_{k=1}^{l} p_{ik} \right) = l \sum_{i=1}^{\hat{I}} b_i + \sum_{i=1}^{\hat{I}} \sum_{k=1}^{l} p_{ik}; \tag{3-1}$$

$$V_\alpha = \sum_{i=1}^{\hat{I}} \left(b_i + \sum_{k=1}^{l} p_{ik} \right) = \sum_{i=1}^{\hat{I}} b_i + \sum_{i=1}^{\hat{I}} \sum_{k=1}^{l} p_{ik}, \tag{3-2}$$

$$V_\iota = \sum_{i=1}^{j} b_i + \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{k=1}^{l} p_{nmk}. \tag{3-3}$$

Обозначим

$$B = \sum_{i=1}^{\hat{I}} b_i, \tag{3-4}$$

$$F = \sum_{i=1}^{\hat{I}} \sum_{k=1}^{l} p_{ik}, \tag{3-5}$$

$$F' = \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{k=1}^{l} p_{nmk}. \tag{3-6}$$

В этом случае выражения (3-1), (3-2), (3-3) преобразуются к виду:

$$V_\sigma = l \cdot B + F, \tag{3-1а}$$

$$V_\alpha = B + F, \tag{3-2а}$$

$$V_\iota = B + F'. \tag{3-3а}$$

Сравним попарно объемы V_σ с V_α и V_α с V_ι. При этом можно сделать следующие выводы.

1. Очевидно, что объем словаря α-типа меньше объема словаря σ-типа на Δ' символов, где

$$\Delta' = (l-1) \cdot B. \tag{3-7}$$

Причем Δ' возрастает с ростом числа словоформ каждого слова словаря и с увеличением количества слов в словаре.

2. *Объем словаря ι-типа отличается* от объема словаря α-типа на величину Δ'', где

$$\Delta'' = F - F'. \tag{3-8}$$

Из (3—5) следует, что F возрастает с увеличением количества лексем в словаре и числа словоформ каждой лексемы пропорционально. Функция F' носит экспоненциальный характер. Это вытекает, во-первых, из того очевидного факта, что чем меньше слов содержится в словаре ι-типа, тем меньшее число парадигм окончаний каждого из грамматических классов будет представлено в массиве типовых парадигм. Во-вторых, существует предел $\lim_{j \to \infty} F'(\iota)$, и этот предел (асимптота $F'(\iota)$) равен F'_{max}, где

$$F'_{max} = \sum_{n=1}^{N_{max}} \sum_{m=1}^{M_{max}} \sum_{k=1}^{l} (p_{nmk})_{max}. \tag{3-6а}$$

F'_{max} представляет собой объем максимального набора типовых парадигм ($M = M_{max}$) всех грамматических классов слов ($N = N_{max}$) при максимальной длине для парадигмы каждого грамматического класса $p_{nmk} = (p_{nmk})_{max}$. Далее (см. 3.3) будет определено, что $N_{max} = 3$, остальные максимальные значения параметров, определяющие F'_{max}, сведены в таблицу 3.1.

Таблица 3.1

Максимальные значения параметров, определяющих типовые парадигмы окончаний

Грамматические классы	*Число типовых* парадигм	*Длина поля* в парадигме	*Число полей* одной парадигмы
Субстантивы	190	5	12
Глагольные слова	328	7	33
Адъективы	31	6	32

Из сказанного ясно, что функция F' имеет следующий вид

$$F' = F'_{max}(1 - e^{-\chi \cdot \iota}), \tag{3-6б}$$

где коэффициент χ определяется экспериментальным путем.

Из выражений (3—5), (3—6б) и (3—8) следует, что объем словаря α-типа будет меньше, чем объем словаря ι-типа при всех $i \leqslant \hat{I}$ в случае, если существует такое $i = \hat{I}'$ (см. рис. 3.3а), что

$$F(\hat{I}) = F'(\hat{I}'). \tag{3-9}$$

В противном случае всегда $\Delta'' > 0$ (см. рис. 3.3б), т. е. объем словаря α-типа больше объема словаря ι-типа, причем разница в объемах увеличивается с ростом количества слов в словарях.

Графическая зависимость объемов словарей α-типа, ι-типа и σ-типа от количества слов в словаре для подъязыка вычислительной техники представлена на рис. 3.4. При этом приняты следующие допущения, полученные в результате лингвостатистических исследований:

1) средняя длина русской машинной основы $b_{icp} = 10$ символам;

2) средняя длина русского машинного окончания $p_{kcp} = 5$ символам;

3) среднее число словоформ одного русского слова $l=10$;

4) слова включались в словарь в порядке убывания частоты их встречаемости в тексте объемом 250 000 словоформ. На графике (рис. 3.4) представлены зависимости объемов словарей всех трех типов от количества слов в словаре, причем для словаря t-типа даны две зависимости:

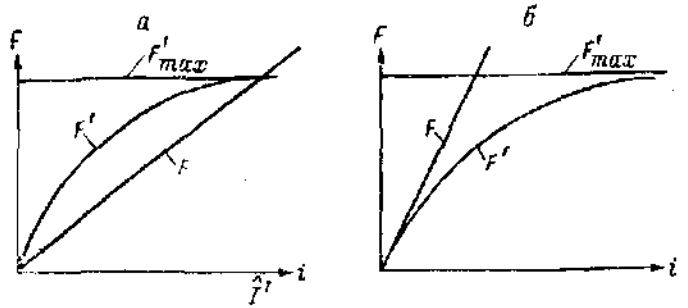


Рис. 3.3. Различные случаи зависимости объемов F и F' от количества слов в словарях.

1) максимально возможный объем словаря $V_{t\max}$, который будет иметь место в случае наличия полного набора типовых парадигм окончаний независимо от числа слов в словаре, т. е. когда $F=F_{max}$;

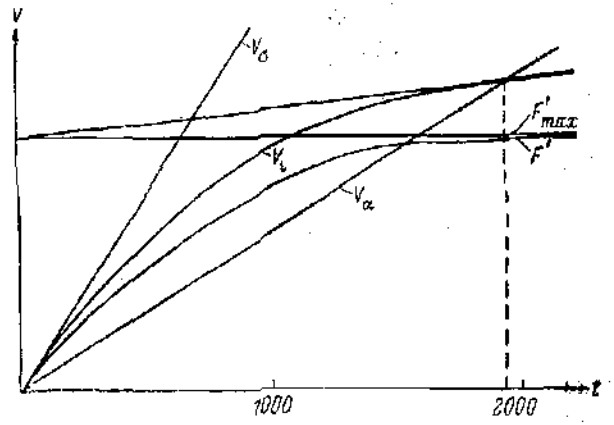


Рис. 3.4. Графическая зависимость объема русских словарей α-, s- и t-типа от числа основ в словаре.

2) действительный объем словаря для $F'=F'_{\max}(1-e^{-\chi i})$. Из сравнения зависимостей $V_{t\max}=V_{t\max}(i)$ и $V_t=V_t(i)$ видно, что они совпадают с точностью не менее 5%, начиная с числа слов в словаре $\geqslant 2000$. Так как практически это всегда имеет место даже для любого узкоотраслевого словаря, все дальнейшие рассуждения будем проводить исходя из максимально возможного объема

$V_t=V_{t\max}$. На основании проведенного анализа объемов словарей σ-, α- и t-типов можно сделать следующие выводы.

1. Словарь словоформ σ-типа по объему всегда, независимо от количества слов в словаре, больше, чем любой словарь основ. Но при этом разница в объеме не постоянна, а отношения объемов $K_{\sigma t}=V_\sigma/V_t$ и $K_{\alpha t}=V_\alpha/V_t$ носят экспоненциальный характер (см. рис. 3.5). Преимущества изолирующей машинной морфологии в словаре σ-типа очень незначительны при реализации информационной базы в структуре банка данных, где могут быть отражены

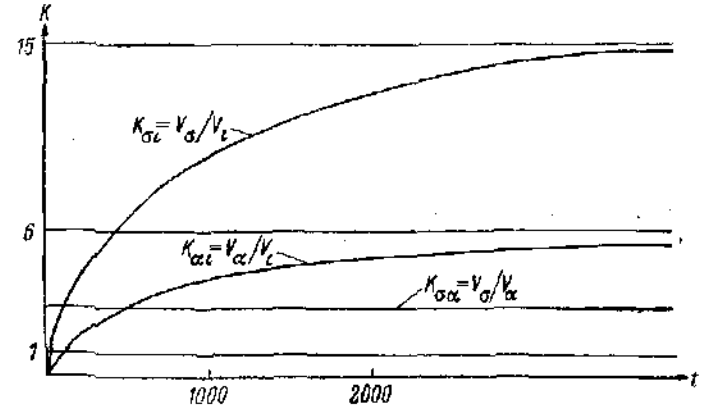


Рис. 3.5. Графическое представление эффективности использования словарей различных типов в зависимости от числа основ в словаре.

все логические связи между отдельными элементами АС и автоматизированы функции их совместной обработки.

2. Словарь основ α-типа, в котором используется сжатие по методу Купера, имеет меньший объем, нежели словарь словоформ σ-типа.

3. Словарь основ t-типа имеет объем меньший, чем словарь α-типа, при количестве слов в словаре, превышающем 2000 (см. рис. 3.4), так как, согласно выражению (3—9), $\Delta''>0$ при $i>\hat{I}'$.

На основании вышесказанного приходим к заключению, что словарь основ агглютинативного типа является наиболее удобной формой организации информационной базы универсальных АС по критерию минимального объема и относительной простоте машинной морфологии.

3.3. Элементная структура автоматических словарей

Каждая единица словаря — его словарная статья — должна содержать, кроме лексической единицы, лексико-грамматическую информацию о ней, признак фразеологичности, признак терминологичности (отнесенности к определенному подъязыку), семанти-

ческий код слова, тезаурусную информацию, т. е. указывать на родо-видовые, ассоциативные, синонимические и антонимические связи слова. Вся эта информация необходима для выполнения различных алгоритмов автоматической переработки текста. Именно введение всех этих сведений в АС даст возможность использовать их как универсальную информационную базу ЛО. Определим элементарную структуру АС, считая, что элементы автоматических словарей — это информация словарных статей, представленная в виде отдельных множеств. Для этого рассмотрим перечисленные виды информации ССт более подробно.

Лексическая единица иноязычного и русского словников была определена в 3.2: для иноязычного словника — это словоформа, для русского — машинная основа. В русском словаре наряду с множеством основ должно быть задано множество машинных окончаний.

Далее выделяются следующие множества: множество грамматических кодов, множество семантических кодов, множество подъязыков семантического пространства, множество фразеологизмов (словарь оборотов). Каждое из них может быть охарактеризовано следующим образом.

Грамматическая и лексико-грамматическая информация. Кодирование грамматической и лексико-грамматической информации для каждой ЛЕ словаря производится согласно стандартной схеме с помощью специальной кодировочной таблицы, содержащей 17 категорий (Борисевич и др., 1970).

Таблица 3.2

Основные грамматические классы, используемые в системе АС

Изменяемые части речи			Неизменяемые части речи
Субстантивы	Адъективы	Глагольные слова	
1. Существительные 2. Количественные числительные 3. Обобщенно-предметные местоимения	1. Прилагательные 2. Порядковые числительные 3. Обобщенно-качественные местоимения	1. Глаголы 2. Деепричастия 3. Причастия	1. Союзы 2. Предлоги 3. Наречия 4. Частицы

Схема (см. табл. 3.2) включает 9 грамматических классов — существительные, прилагательные, числительные, местоимения, глаголы (вместе с деепричастиями и причастиями, которые склоняются по моделям, предусмотренным для прилагательных), наречия, предлоги, союзы и частицы. Среди этих классов первые пять включают изменяемые формы, а последние четыре дают неизменяемые слова. Поскольку количественные числительные и

обобщенно-предметные местоимения изменяются как существительные, а порядковые числительные, обобщенно-качественные местоимения склоняются как прилагательные, все знаменательные части речи можно сгруппировать в три подмножества: подмножество субстантивов; подмножество адъективов; подмножество глагольных слов.

Схема кодирования лексико-грамматической информации может быть представлена в виде некоторой матрицы вида:

$$C=\begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1j} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2j} & \cdots & g_{2n} \\ \vdots & & & & & \\ g_{i1} & g_{i2} & \cdots & g_{ij} & \cdots & g_{in} \\ \vdots & & & & & \\ g_{m1} & g_{m2} & \cdots & g_{mj} & \cdots & g_{mn} \end{bmatrix},$$

где каждая i-тая строка представляет совокупность кодов лексико-грамматических категорий (признаков), соответствующую каждой единице словаря. В матрице допускается наличие нулевых (незаполненных) строк, которые могут быть использованы при возможных пополнениях АС. Дополним каждую строку матрицы порядковым номером строки d_i. Матрица приобретает вид:

$$G=\begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1j} & \cdots & g_{1n} & d_1 \\ g_{21} & g_{22} & \cdots & g_{2j} & \cdots & g_{2n} & d_2 \\ \vdots & & & & & & \\ g_{i1} & g_{i2} & \cdots & g_{ij} & \cdots & g_{in} & d_i \\ \vdots & & & & & & \\ g_{m1} & g_{m2} & \cdots & g_{mj} & \cdots & g_{mn} & d_m \end{bmatrix}.$$

При каждой словарной статье АС можно указывать не всю строку кодов лексико-грамматической информации, а только порядковый номер строки d_i, так называемый сжатый код (СК), т. е. информацию, содержащуюся в матрице D вида:

$$D=\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_i \\ \vdots \\ d_m \end{bmatrix}.$$

Богатство формообразования в языке, особенно русском, приводит к тому, что обычно одно словообразовательное гнездо в ма-

трицах C и D представлено несколькими строками, т. е. подмножеством $\{d_i\} \in D$, где $i=\overline{1, \tilde{m}}$, $\tilde{m}<m$. Это подмножество в свою очередь можно заменить одним кодом, так называемым инициальным сжатым кодом (ИСК), который характеризует парадигму лексической единицы.

Множество ИСК образуют матрицу вида:

$$K=\begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_j \\ \vdots \\ k_l \end{bmatrix}, \text{ где } l<m.$$

В целях уменьшения объема словаря лексико-грамматическую информацию словарных статей, представленную элементами матриц C, G, D и K, целесообразно вынести в отдельные множества. Это объясняется тем, что многие словоформы языка — как русские, так и иноязычные — дают значительный процент грамматической омографии; при организации отдельных множеств C, G, D и K омография устраняется. Кроме того, такая организация удобна и с точки зрения использования лексико-грамматической информации в различных ЛЗ. Громоздкие кодировочные таблицы C и G требуются при выполнении далеко не всех лингвистических алгоритмов; грамматический анализ можно проводить на уровне сжатых кодов или инициальных кодов, т. е. пользоваться только значениями $d \in D$ или $k \in K$.

Следует отметить, что начальная матрица C далее вообще не включается в элементную структуру словаря, так как вся ее информация полностью содержится в матрице G.

Семантическая информация. Использование МАРС в качестве универсального хранилища сведений, необходимых для решения разнообразных задач, начиная с перевода иноязычного текста и кончая организацией диалога «человек—машина», требует введения в каждую ССт словаря разнообразной семантической информации.

Эта информация охватывает следующие типы сведений: указание на терминологичность или нетерминологичность слова, семантическую характеристику слова, указание на связи понятия, выражаемого словом, с другими понятиями.

Задание сведений (кодов) первого типа потребовало разделить все семантическое надпространство, охватываемое языком, на предметные области — частные семантические пространства (СП), реализуемые в тематически ограниченных совокупностях текстов, так называемых подъязыках.

Семантическая характеристика ССт включает указание на принадлежность слова к одному определенному подъязыку (в этом

случае слово является терминологичным), либо двум и более подъязыкам с указанием их кодов (для политерминологических слов), либо, наконец, указание на употребительность слова во всех подъязыках (общеупотребительное слово).

Нельзя забывать также о том, что существует большое число терминов, употребляющихся в одном и том же или близких значениях в разных подъязыках (например, *таблица* и *матрица*). Обычно в значении такого термина выделяется некоторый семантический центр, определяющий его принадлежность к основному подъязыку, а также дополнительные семантические характеристики, которые указывают на принадлежность данного слова к другим предметным областям, для которых центральными являются другие слова.

Код терминологичности используется как при решении задач распознавания общего смысла, так и при решении задач снятия многозначности отдельных ЛЕ в ходе МП иноязычного документа.

В первом случае этот код служит указанием на то, что слово является ключевым или полиключевым в данной предметной области и может входить в эталоны, использующиеся при автоматической атрибуции, аннотировании и реферировании текста.

Во втором случае код терминологичности, относя слово к определенному подъязыку, одновременно указывает и на его точное значение. В качестве примера укажем на ЛЕ *слово*, которое, будучи соотнесенным с подъязыком вычислительной техники, приобретает значение «поле основной памяти ЭВМ с фиксированной длиной в 4 байта» (Единая система ЭВМ, 1974, с. 6).

Напротив, перенесенное в подъязык языкознания, оно имеет значение «предельная составляющая предложения, способная непосредственно соотноситься с предметом мысли как обобщенным отражением данного участка действительности и направляться на эту последнюю. . .» (Ахманова, 1966, с. 422).

Семантический код слова предусматривает отнесение его к определенной лексико-семантической группе. Этот код вырабатывается относительно каждого подъязыка с помощью классификационных древовидных графов, построенных на использовании лингвистических и экстралингвистических (логических, психологических и натурно-физических) признаков. В тех случаях, когда неоднозначность ЛЕ не может быть устранена с помощью словаря оборотов или кода терминологичности, эта задача решается с помощью специальных алгоритмов, исследующих сочетаемость семантических кодов.

Тезаурусная информация. Более или менее полное извлечение с помощью ЭВМ смыслового инварианта документа и определение его ремы (новизны), а также организация диалога «человек — машина» становятся возможными только тогда, когда в машину вводится структурированное тезаурусное описание всего семантического пространства или его отдельных

областей. *В нашем случае в роли тезауруса выступает парадигматическая модель плана содержания, представляющая собой* множество лексических единиц, между которыми заданы смысловые связи как иерархического (родо-видовые связи), так и неиерархического (синонимия, антонимия, ассоциация) типов.

Принадлежность данного слова к тезаурусной модели, а также его связи отмечаются в ССт (см. рис. 1.2). Вся совокупность тезаурусных связей слова с другими словами, а также формы представления этих связей показаны в таблице 3.3. Следует подчер-

Таблица 3.3

Типы тезаурусных отношений в системе АС

Родо-видовые отношения слова		Ассоциативные отношения слова		
Связь с родовыми терминами	Связь с видовыми терминами	Связь с синонимическими терминами	Связь с антонимическими терминами	Связь с ассоциативными терминами с указанием характера ассоциации

кнуть, что тезаурусная информация в АС задается связями различных типов, а не определяется отдельными элементами (множествами) структуры словарей, как это имело место относительно других видов информации ССт.

Множества, выделенные в результате анализа информации ССт автоматических словарей, идентифицируются как элементы структурируемой информационной базы $\{e_i\}$. Их параметры представлены в таблице 3.4. Физически каждый экземпляр элемента АС размещается в поле данных информационной базы. Множество полей данных $Q=\{q_m, p_m\}$ и их характеристик определяется на данном этапе, несмотря на то что поле — параметр физической структуры базы. Это объясняется тем, что параметры поля — множество элементов $\{e_i\}_m$, длина поля p_m и формат представления данных в поле f_m^q — зависят от элементной структуры словаря и требований минимизации объема базы и времени процедур обработки данных БОЛИД, которые определены в 3.1 и 2.3 соответственно.

При размещении элементов данных в поля в этой связи используются следующие принципы:

1) каждому элементу данных отводится отдельное поле;

2) те элементы данных, которые при их обработке принимают участие в арифметических операциях, представлены в двоичном формате, остальные элементы используются в символьном формате. Характеристики полей данных информационной базы АС представлены в таблице 3.4.

Таблица 3.4

Основные характеристики элементов информационной базы АС

Элемент АС	Множественное представление элементов АС	Количество значений элементов в АС	Длина поля элемента АС в байтах	Формат представления данных в поле элемента АС	Примечания
Подъязык	Q	43	1	Двоичный	Не зависит от типа подъязыка и количества подъязыков в АС
Развернутый грамматический код	G	916	15	»	То же
Сжатый грамматический код	D	916	4	»	»
Инициальный грамматический код	K	85	4	»	»
Сжатый грамматический код иноязычного АС	N	100		»	»
Семантический код	S	—	2	»	»
Русская машинная основа	B	10000	29	Символьный	Для системы АС по вычислительной технике
Набор русских машинных окончаний	F	655	231	»	Не зависит от типа подъязыка и количества подъязыков в АС
Русский словарь оборотов	O	—	65	»	Для системы АС по вычислительной технике
Английское слово	M	13000	23	»	То же
Иноязычный словарь оборотов	Z	—	55	»	»

3.4. Выводы

1. Для формирования базы лингвистических данных должны быть определены следующие задачи структурирования: множество элементов автоматических словарей и единая информационная модель (информационный граф) системы АС.

2. Элементная структура АС определяется по критерию минимизации объема информации словарей, так как даже по предварительной оценке системы многоотраслевого русского АС объем последнего превышает 350 млн. символов.

3. Уменьшение объема АС производится за счет использования в качестве русского АС словаря машинных основ, а в качестве английского АС (для случая двуязычной ситуации) словаря словоформ.

4. Лексико-грамматическая, семантическая, тезаурусная, фразеологическая характеристики ЛЕ словарей выделяются в отдельные множества элементов с целью исключения дублирования этой информации в различных ССт АС.

Заключение

В результате проведенных исследований определена общая методика поэтапного синтеза логических и физических структур информационной базы основных лингвистических данных. Основное внимание в работе было уделено лингвистическим аспектам структурирования БОЛИД — этапу формирования элементной структуры системы автоматических словарей. Синтез элементной структуры АС производился по критерию минимизации объема информационной базы БОЛИД.

Л. В. Малаховский

СТРУКТУРНЫЕ И КВАНТИТАТИВНЫЕ ХАРАКТЕРИСТИКИ ОМОНИМИЧЕСКИХ РЯДОВ В СОВРЕМЕННОМ АНГЛИЙСКОМ ЯЗЫКЕ

I. Структура омонимического ряда

1. Предварительные замечания

Ни одно точное описание словарного состава языка не может обойтись без глубокого рассмотрения вопросов омонимии. Не меньшее значение имеет разработка проблем омонимии для решения задач инженерной лингвистики.

Омонимия есть проявление того особого свойства языкового знака, благодаря которому разным означаемым могут соответствовать тождественные означающие. В силу этого омонимия представляет собой определенную помеху в процессе коммуникации. Часто слушающий оказывается в затруднении, какое из нескольких разных значений, выражаемых данной языковой формой, следует выбрать для правильного понимания сообщения. Затруднения могут возникать не только у слушающего (читающего), как это обычно принято думать (см., например: Ахманова, 1968, с. 8), но и у говорящего (пишущего), который в целях оптимизации процесса общения стремится строить высказывание так, чтобы оно могло быть понято однозначно. Омонимия вносит трудности и в процесс усвоения иностранного языка, когда учащийся сталкивается с тем, что одна и та же языковая форма может иметь совершенно разные значения — факт, на который в своем родном языке он обычно не обращает внимания. Анализ таких форм существенно затрудняет восприятие иноязычного текста или сообщения.

Столь же значительные трудности вызывает омонимия и при построении систем автоматической переработки текста, а также в лексикографической работе: при выборе единиц счета для частотного словаря, при определении границ между отдельными сло-

вами в общем словаре языка, при разработке способов представления в нем омонимичных слов и т. п. Проблемы омонимии все более серьезно обсуждаются также в связи с широко развернувшейся работой по статистическому изучению лексики (см.: Алексеев, 1971, с. 299—301, а также Шайкевич, 1962, с. 271 сл., Ullmann, 1964, с. 79, Allén, 1970—1971) и в связи с использованием омонимов в экспериментальных психолингвистических и психологических исследованиях (Cramer, 1970, с. 361; Leah, 1972, с. 255; и др.)

Работа по созданию эффективных программ АПТ, по составлению и совершенствованию толковых и переводных словарей, по улучшению методики преподавания языков, как иностранных, так и родного — все это настоятельно требует углубленного исследования проблем омонимии, выявления ее внутренних закономерностей, определения места, которое она занимает в общей системе языка.

К сожалению, несмотря на то что изучение омонимии ведется уже давно, сделано в этой области в силу ряда причин еще очень мало. Долгое время омонимия исследовалась по преимуществу не в синхронном, а в диахроническом плане. Основное внимание уделялось вопросам возникновения, борьбы и взаимного вытеснения омонимов (ср.: Gilliéron, 1921; Булаховский, 1928, с. 47; Öhmann, 1934; с. 1; Menner, 1936, с. 229; Kieft, 1938; Williams, 1944; и др.). Синхронное изучение омонимии сводилось почти исключительно к поискам критерия разграничения омонимии и полисемии (обзор литературы по этому вопросу см. в работе: Задорожный, 1971), что, при недостаточной разработанности семасиологии и полной неизученности структурных характеристик омонимических рядов в конкретных языках, не могло привести к успешным результатам.

К настоящему времени омонимия оказалась изученной значительно слабее, чем такие смежные явления, как полисемия, синонимия или антонимия. До сих пор нет четкого определения основных понятий омонимии. Более того, многие важнейшие понятия омонимии, как будет показано ниже, даже и не выявлены вообще. В большой степени не упорядочена терминология, в результате чего одно и то же понятие обозначается разными авторами по-разному, и, наоборот, для обозначения разных понятий нередко используются одни и те же термины (см.: Малаховский, 1973, с. 47—49). Нет классификации, которая выявляла бы все возможные виды омонимических единиц данного языка и адекватно отражала бы характер формально-смысловых отношений между ними.[1]

Отсутствие точного, строго последовательного описания омонимии в плане парадигматики мешает осуществлению ее синтагматического описания, в котором так нуждаются инженерная лингвистика и методика преподавания языков. Задача настоящего исследования и состоит в том, чтобы дать такое парадигматическое описание и на его основе выявить некоторые, наиболее существенные, количественные характеристики омонимических рядов.

[1] Классификация, разработанная нами ранее (Малаховский, 1973, с. 50—54), учитывала только омонимию на уровне слов и не рассматривала омонимию на уровне словоформ.

2. Основные понятия омонимии и принципы выделения омонимов

Как было показано ранее (Малаховский, 1975), омонимический ряд представляет собой своего рода микросистему, целостность которой базируется на общности формы ее элементов. Изучение омонимического ряда может идти как по линии исследования его системных свойств (чему и была посвящена упомянутая работа), так и по линии анализа его структуры. В первом случае омонимический ряд рассматривается как сложное единство, и тогда нас интересуют его свойства как целого, во втором же он рассматривается как расчлененное множество, и тогда нас интересуют его состав и внутренняя организация. Именно этот второй подход и используется в настоящей работе.

Работа строится на материале современного английского языка, однако отдельные ее выводы могут, как нам кажется, иметь и универсальное значение. Следует подчеркнуть, что в работе рассматривается только омонимия слов и синтетических словоформ.

Обратимся к рассмотрению основных понятий омонимии.

Основным структурным элементом омонимики языка является о м о н и м и ч е с к и й р я д. Под омонимическим рядом понимается совокупность слов (лексем) или словоформ одного и того же языка в один и тот же период его существования, тождественных друг другу в плане выражения, но существенно различающихся между собой в плане содержания.

Омонимика языка представляет собой совокупность всех омонимических рядов языка, одни из которых находятся между собой в определенных системных отношениях (например, ряды исходных и производных омонимов или ряды основных и косвенных омоформ), другие же никак между собой не связаны. Таким образом, омонимика в целом единой системы не образует. В этом отношении она подобна синонимике (см.: Смирницкий, 1956, с. 204).

Важно различать два вида омонимических рядов: ряды, состоящие из омонимичных друг другу слов — омонимов, и ряды, образованные омонимичными формами слов — омоформами. Ряды первого вида мы будем называть омонимическими группами, или о м о г р у п п а м и (ОГ), а ряды второго вида — омонимическими комплектами, или о м о к о м п л е к т а м и (ОК).

Выделение омонимов в том или ином конкретном языке, т. е. установление наличия омонимических отношений между словами этого языка, представляет большие трудности. Выявление факта

тождества слов в плане выражения затрудняется тем, что слово может выступать не в одной, а в нескольких разных грамматических формах. Поэтому характер формального тождества слов может быть различным: слова могут совпадать во всех своих формах или только в некоторых. Если мы хотим, чтобы определение омонимов охватывало оба эти случая, то необходимо условиться, что для признания двух слов фонетически тождественными достаточно, чтобы хотя бы одна из форм одного слова совпадала хотя бы с одной из форм другого. Отсюда ясно, что для решения вопроса о формальном тождестве (или различии) двух слов необходимо сопоставить все их грамматические формы. Иначе говоря, вопрос о формальном тождестве слов решается не на уровне слов, а на уровне словоформ.

Исследование формально-смысловых отношений на уровне словоформ, т. е. не между словами в целом, а между отдельными формами слов, позволяет судить о том, в каких случаях омонимия является п о л н о й (совпадают попарно все формы сравниваемых слов, например: ear, ears 'ухо' и ear, ears 'колос'), в каких — ч а с т и ч н о й (совпадает лишь часть парадигмы одного слова с частью парадигмы другого, die, dies 'штамп' и die, dice 'игральная кость'), а в каких — н е р а в н о о б ъ е м н о й [2] (все формы одного слова совпадают с соответствующими формами второго, тогда как для некоторых форм второго не находится соответствующих форм у первого: chink, chinks 'щель' и chink, chinks, chinking, chinked 'звякать').

Сопоставление всех — основных (исходных, словарных) и косвенных — форм омонимичных слов дает возможность выявить не только такие омонимы, которые совпадают в своих исходных формах (так называемая с л о в а р н а я омонимия), но и такие, у которых совпадают лишь косвенные формы, например, butted *pt* и *pp* от BUTT *v* 'бодаться' и butted *pt* и *pp* от BUT *v* 'возражать' (н е с л о в а р н а я омонимия), или словарная форма одного совпадает с косвенной формой другого, например, RUNG *n* 'ступенька' и rung *pp* от RING *v* 'звенеть' (с м е ш а н н а я омонимия).[3]

При установлении формального тождества слов необходимо учитывать также и явления фонетического и графического варьирования. Слово может иметь два, а иногда и больше, фонетических, графических или фонетико-графических варианта (например: PEDLAR/PEDLER ['pedlə] *n* 'разносчик', POSTPONE [poust'poun/pous'poun] *v* 'откладывать', THRASH [θraeʃ]/THRESH [θreʃ] 'бить'), причем в каждом из этих вариантов слово (если оно является изменяемым) обладает полной системой грамматических форм. Могут варьировать и отдельные грамматические формы

[2] Термин Ю. С. Маслова (Маслов, 1963, с. 199).

[3] Здесь и далее прописными буквами даются слова (лексемы), а строчными — словоформы. Слова или словоформы, выступающие в качестве названий омогрупп или омокомплектов, выделяются полужирным шрифтом.

слова (например: ate [eɪt/et] *pt* от EAT *v* 'есть', spoiled [spoild] / spoilt [spɔilt] *pt* и *pp* от SPOIL *v* 'портить'). Поэтому для выявления формального тождества двух слов необходимо сопоставление не только всех их грамматических форм, но и всех фонетических и графических вариантов, допускаемых литературной нормой (см.: Allén, 1970, т. I, с. XXXV).

Для признания двух слов формально тождественными достаточно, чтобы хотя бы один из вариантов одного слова был тождествен хотя бы одному из вариантов другого. В соответствии с этим слова, выступающие в двух или нескольких вариантах, следует считать омонимами как в случае, когда они тождественны друг другу во всех своих вариантах (ср.: $SEAR_1$/$SERE_1$ *a* 'сухой', 'увядший' и $SEAR_2$/$SERE_2$ 'спусковой рычаг'), так и тогда, когда совпадают лишь некоторые из их вариантов (как, например, $CODLING_1$ *n* 'мелкая треска' и $CODLING_2$/CODLIN *n* 'сорт яблок').

Еще бо́льшие затруднения вызывает установление факта смыслового различия слов, т. е. различия их в плане содержания. Это объясняется тем, что не всякое различие значений двух формально тождественных языковых единиц является признаком их омонимичности. Ситуация, при которой одному и тому же звучанию (написанию) соответствует несколько разных значений, встречается в языке очень часто, однако в одних случаях эти разные значения рассматриваются как относящиеся к разным словам (омонимия), а в других — как принадлежащие одному и тому же слову (полисемия).

Вопрос о принадлежности двух формально тождественных, но различающихся по значению единиц одному и тому же слову или разным словам решается по-разному, в зависимости от того, какие значения имеются в виду — грамматические или лексические.

При рассмотрении грамматических значений необходимо, прежде всего, разграничивать значения в н у т р е н н и е, т. е. специфичные для отдельных грамматических форм одного и того же слова (как, например, значения числа и падежа у имени существительного, значения лица, времени и др. у глагола), и в н е ш н и е (категориальные), присущие всей системе форм слова в целом (например, значение части речи или — в языках, имеющих категорию грамматического рода — значение рода у имен существительных).[4] Различие внутренних грамматических значений использоваться в качестве признака омонимии (на уровне слов) не может, потому что языковые единицы, различающиеся по своим внутренним значениям, могут принадлежать как разным словам (например, $fish_1$ *n sg* 'рыба' — $fish_2$ *n sg* 'фишка'), так и одному и тому же слову ($fish_1$ *n sg* 'рыба' — $fish_1$ *n pl* 'рыбы'). Наоборот, различие внешних грамматических значений является сущест-

[4] См.: Кацнельсон, 1972, с. 23.

венным. Оно всегда указывает на омонимию, так как разные категориальные значения в рамках одного слова несовместимы: слово не может принадлежать двум разным частям речи одновременно (см.: Смирницкий, 1956, с. 74—75).

При рассмотрении лексических значений признаком омонимичности двух языковых единиц, тождественных в плане выражения, является отсутствие смысловой связи между их означаемыми, устанавливаемое на основе компонентного анализа или путем сравнения соответствующих словарных толкований. Если в толкованиях (семантических деревьях) двух означаемых или в соответствующих наборах семантических компонентов нет общей части или имеется лишь тривиальная общая часть (ср.: Апресян, 1974, с. 185), то такие означаемые по смыслу не связаны. В этом случае различия в лексических значениях можно считать существенными и сравниваемые языковые единицы рассматривать как разные слова — омонимы; в противном случае мы будем говорить о тождестве лексических значений.

На основе сказанного сформулируем более точное и полное определение омонимов. Омонимы — это слова одного и того же языка в один и тот же период его существования, тождественные друг другу фонетически и/или графически во всех или некоторых своих грамматических формах (и во всех или некоторых фонетических или графических вариантах), но существенно различающиеся по лексическим и/или грамматическим значениям.

3. *Классификация формально-смысловых отношений между омонимичными словами*

Традиционная лингвистика делит омонимы на омофоны (омонимы, тождественные фонетически, но различающиеся графически), омографы (омонимы, тождественные по письменной форме, но различающиеся по звуковой) и «полные» омонимы (совпадающие как по звучанию, так и по написанию). Для некоторых специальных целей (составление словарей и учебных пособий по орфографии и орфоэпии, отбор языкового материала для психолингвистических исследований и т. п.) выделение таких категорий оправдано, однако для анализа конкретного речевого (текстового) материала эта классификация непригодна.

Дело в том, что при восприятии звуковой речи мы сталкиваемся только с такими омонимами, которые совпадают по звучанию; при этом нам неизвестно, совпадают они также и на письме или нет. Воспринимая письменную речь, мы можем встретить только такие омонимы, которые совпадают по написанию; при этом в письменном представлении этих слов нет никаких указаний на то, совпадают ли они также и по звучанию. Таким образом, актуально значимым является деление омонимов не на три типа, а на два: ф о н е т и ч е с к и е — тождественные по звучанию (независимо от соотношения их письменной формы) и г р а ф и ч е с к и е — тождественные по написанию (независимо от соотношения их звуковой формы). Тип фонетических омонимов охватывает два типа традиционной классификации — омофоны и «полные» омонимы, а тип графических омонимов — тоже два: омографы и те же «полные» омонимы. Отсюда видно, что названные два типа омонимов являются пересекающимися: в качестве их общей части выступают «полные» (точнее — фонетико-графические) омонимы, которые потому и попадают в оба эти типа, что являются омонимами как по звучанию, так и на письме.

Исходя из того, что наше исследование ориентировано главным образом на интересы инженерной лингвистики и лексикографии, т. е. на интересы отраслей, имеющих дело почти исключительно с письменным вариантом языка, мы будем рассматривать здесь только графические омонимы (называя их просто омонимами); отметим при этом, что используемый нами подход полностью применим и для описания омонимов фонетических.

Рассматривая омонимы в плане соотношения лексических и грамматических значений, мы обнаруживаем также два типа омонимов — омонимы л е к с и ч е с к и е, т. е. различающиеся между собой по «собственно лексическим» (Кацнельсон, 1972, с. 89) компонентам значения, и омонимы г р а м м а т и ч е с к и е — различающиеся по грамматическим значениям, в первую очередь по категориальному значению части речи.[5] Поскольку при анализе конкретных омонимических пар нам придется учитывать сразу оба эти признака — соотношение (т. е. тождество или различие) лексических значений и соотношение значений грамматических, мы фактически будем иметь дело не с двумя названными типами омонимов, а с подтипами, образующимися в результате их сопоставления (пересечения). Таких подтипов три:

1) омонимы ч и с т о л е к с и ч е с к и е — различающиеся по лексическим и тождественные по грамматическим значениям, например: DIE_1 *v* ‘умирать’ — DIE_2 *v* ‘штамповать’ или BOW_1 [bau] *n* ‘поклон’ — BOW_2 [bou] *n* ‘лук’, ‘дуга’;

2) омонимы ч и с т о г р а м м а т и ч е с к и е — различающиеся по грамматическим и тождественные по лексическим значениям, например: FALL *v* ‘падать’ — FALL *n* ‘падение’ или USE [ju:s] *n* ‘польза’ — USE [ju:z] *v* ‘использовать’;

3) омонимы л е к с и к о - г р а м м а т и ч е с к и е, различающиеся по тем и другим значениям одновременно, например: $CONTINENT_1$ *n* ‘материк’ — $CONTINENT_2$ *a* ‘сдержанный’ или TEAR [tɪə] *n* ‘слеза’ — TEAR [tɛə] *v* ‘рвать’.

Термины *«чисто лексические» и «лексико-грамматические»* не являются новыми, они встречаются, в частности, и у А. И. Смир-

[5] Мы намеренно отвлекаемся здесь от весьма сложной проблемы разграничения лексических и грамматических значений (см.: Кацнельсон, 1972, с. 89—93), поскольку номенклатура выделяемых типов омонимов от ее решения не зависит.

ницкого (Смирницкий, 1956, с. 166). В пояснениях нуждается лишь термин «чисто грамматические».

Необходимо сразу же сделать оговорку, что уточнитель «чисто» здесь вовсе не означает, что между омонимами, к которым этот термин применяется, нет никаких лексико-семантических различий. Различия между лексическими значениями «чисто грамматических» омонимов есть и даже обязательно должны быть: омонимы эти всегда относятся к разным частям речи, а различие между частями речи уже само по себе является различием не только грамматическим, но и лексическим (см.: Смирницкий, 1956, с. 169). Речь идет лишь о том, что вещественный компонент, лежащий в основе лексических значений слов типа FALL *v* 'падать' и FALL *n* 'падение', один и тот же, их словарные толкования обязательно имеют общую часть.[6] Это позволяет нам говорить о тождестве лексических значений и называть подобные омонимы «чисто грамматическими».

Термин «грамматические омонимы», используемый в некоторых работах для обозначения омонимов этого типа (см., например: Костюченко, 1954), недостаточно точен, так как указывает только на различие грамматических значений, но ничего не говорит о тождестве значений лексических. Термин «моделированные омонимы», введенный И. П. Ивановой (Иванова, 1963) сам по себе возражений не вызывает, однако он выпадает из общей системы терминов, поскольку не содержит указаний на характер соотношения лексических и грамматических значений омонимов. Термин «частичная сложная лексико-грамматическая омонимия», используемый А. И. Смирницким, непригоден, поскольку тот же термин применяется в классификации А. И. Смирницкого и для обозначения омонимии типа $BEAR_1$ *v* 'носить' — $BEAR_2$ *n* 'медведь' или $LIGHT_1$ *n* 'свет' — $LIGHT_2$ *a* 'легкий'. Ясно, что омонимы, тождественные по лексическим значениям, и омонимы, не имеющие в своих лексических значениях ничего общего, не должны обозначаться одним и тем же термином. Не спасает положения и термин «грамматико-лексические омонимы», который, по мнению А. Я. Шайкевича, указывает на то, что у омонимов этого типа «различие грамматического значения явно превалирует над различием лексического значения» (Шайкевич, 1962, с. 52). Не говоря уже о том, что вряд ли вообще возможно сравнивать по величине эти два разных вида различий, перестановка компонентов исходного термина «лексико-грамматические» ничего по существу не меняет, поскольку в сложных прилагательных, пишущихся через дефис, оба компонента равноправны. Таким образом, термин «чисто грамматические» (с указанной оговоркой) представляется наиболее точным и полностью укладывающимся в общую систему терминов.

[6] Ср., например, в словаре А. Хорнби и др.: fall *v* 'come down from a higher place'; 'drop down from'; fall *n* 'the act of falling, dropping or coming down' (Hornby et al., 1958, с. 437—438).

Необходимо подчеркнуть, что к чисто грамматическим мы относим не только моделированные омонимы типа FALL *v* — FALL *n*, но и омонимы типа USE *n* 'польза' — USE *v* 'использовать', которые обычно рассматриваются как словообразовательные пары с фонетическим чередованием. С точки зрения звукового варианта языка слова [ju:s] *n* и [ju:z] *v* действительно не являются омонимами; однако поскольку в письменной форме подобных слов никаких указаний на их звучание нет, то в письменном варианте языка пары типа USE *n* — USE *v*, CONTACT *n* 'контакт' — CONTACT *v* 'контактировать' (с переносом ударения) и т. п. ничем не отличаются от омонимических пар типа FALL *n* 'падение' — FALL *v* 'падать', что и дает основание считать их омонимами (графическими) и относить их к подтипу чисто грамматических омонимов. В общеязыковом же плане, т. е. при учете как письменной, так и звуковой формы, подобные пары следует квалифицировать как «чисто грамматические омографы» (ср.: Малаховский, 1973, с. 54).

Некоторые авторы возражают против включения слов, соотносящихся по конверсии, в число омонимов. В. И. Абаев называет такие слова «мнимыми омонимами» и характеризует их как случаи «лексико-синтаксической полисемии» (Абаев, 1957, с. 39). И. С. Тышлер тоже считает, что «конвертированные слова» не являются омонимами, «ибо нет разрыва по линии семантики» (Тышлер, 1966, с. 3—4, 10; см. также: Тышлер, 1975, с. 10).

С таким возражением согласиться нельзя. Совершенно прав был А. И. Смирницкий, утверждая, что «. . . картина английской омонимики, не включающая в свои рамки лексико-грамматические омонимы, связанные с конверсией, является в высшей степени неполной (а потому и неправильной) картиной» (Смирницкий, 1956, с. 173). Причисление слов, находящихся в конверсионных отношениях, к омонимам с неизбежностью вытекает из определения омонимов как слов, различающихся по лексическим и/или грамматическим значениям. Если исключить слова, соотносящиеся по конверсии, из числа омонимов, то придется изменить и само определение омонимов. Определение же это исходит из давно сложившегося в лингвистике понимания омонимов как языковых знаков, имеющих тождественные означающие, но разные означаемые (Bally, 1944/1955, с. 189). Поскольку в понятие означаемого входит не только лексическое, но и грамматическое значение, то ясно, что грамматические, в том числе и чисто грамматические, омонимы имеют такое же право на существование, как и омонимы лексические.

Можно, конечно, попытаться построить описание формально-смысловых отношений между словами так, чтобы термин «омонимия» применялся только к словам, различающимся по лексическим значениям, а чисто грамматическая омонимия образовала особую категорию и стала именоваться как-либо иначе, допустим, «полифункциональностью». Но такое описание не изменило бы

сути дела — того, что «полифункциональность» и «истинная омонимия» представляют собой разновидности некоего более общего явления — различия означаемых при тождестве означающих. Для обозначения этого явления потребовалось бы ввести какой-то новый термин, скажем, «неоднозначность». Но тогда мы вернулись бы к прежнему описанию, с той лишь разницей, что вместо терминов, четко отражающих соотношение обозначаемых категорий (омонимия: а) чисто лексическая, б) лексико-грамматическая, в) чисто грамматическая) использовались бы термины, не отражающие этого соотношения совсем (неоднозначность: а) лексическая омонимия, б) лексико-грамматическая омонимия, в) полифункциональность). Вряд ли такое описание оказалось бы более удобным или более адекватно отражающим описываемые явления. В пользу первого описания говорит и давняя лингвистическая традиция, относящая явление конверсии в английском языке и сходные с ней явления в русском, немецком и других языках к омонимии (см., например: Jespersen, 1928, с. 365; Виноградов, 1940, с. 10—11; Амосова, 1956, с. 85; Ахманова, 1957, с. 164; Левковская, 1956, с. 151—152; Арнольд, 1973, с. 103, Маслова-Лашанская, 1973, с. 54).

4. Анализ структуры омонимического ряда. Уровень слов

Рассмотрим конкретную группу английских омонимов с тождественным написанием **CHUCK**. Эта омогруппа состоит из шести слов:

1) CHUCK *n* ‘цыпленок’;
2) CHUCK *v* ‘кудахтать’; ‘скликать (домашнюю птицу)’;
3) CHUCK *int* ‘цып-цып!’;
4) CHUCK *n* ‘полено’;
5) CHUCK *n* ‘зажимной патрон’;
6) CHUCK *v* ‘зажимать’.[7]

Сопоставляя значения слов, входящих в эту ОГ, видим, что в одних случаях различия между омонимами в плане содержания касаются только лексических значений (например, у слов 1 и 4 или 2 и 6), в других — только грамматических [8] (например, у слов 1 и 3, 5 и 6), в третьих — и тех, и других (1 и 6, 2 и 4 и др.). Из этого следует, что в составе ОГ имеются слова, т о ж д е с т в е н н ы е друг другу по какому-нибудь одному из компонентов

[7] Разграничение омонимов дается по «Большому англо-русскому словарю» под ред. И. Р. Гальперина (М., 1972; далее — БАРС). В целях простоты изложения некоторые из омонимов, приводимые в этом словаре, здесь опущены.

[8] Говоря о различии или тождестве грамматических значений на уровне слов, мы имеем в виду только внешние грамматические значения, т. е. значения частей речи, поскольку различие внутренних грамматических значений для омонимии слов нерелевантно.

означаемого — по лексическому или по грамматическому значению. Так, слова 1, 4 и 5 тождественны по грамматическому значению (все они являются существительными), а слова 1, 2 и 3 — по лексическому значению (все они имеют один и тот же ведущий семантический компонент ‘домашняя птица’).

Таким образом, в составе омогруппы выявляются структурные единицы, более крупные, чем слово. Одни из них объединяют слова, относящиеся к одной и той же части речи, другие — слова, близкие по лексическому значению.

Единицы, охватывающие слова, относящиеся к одной и той же части речи, можно назвать сегментами омогруппы, или о м о-

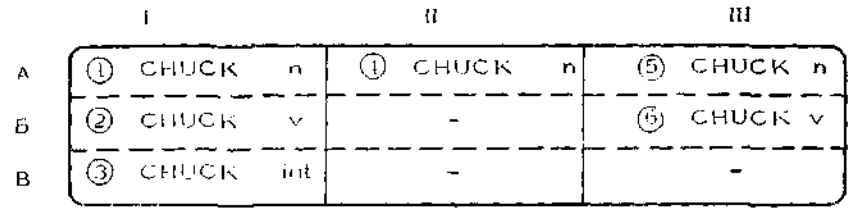


Рис. 1. Членение омогруппы CHUCK на сегменты (А, Б, В) и гиперлексемы (I, II, III). *Цифры* в кружочках соответствуют порядковым номерам слов в исходном списке (стр. 154).

с е г м е н т а м и. Именно омосегмент выступает в качестве единицы счета в частотных словарях, не различающих лексической омонимии, но различающих омонимию грамматическую (см., например: Алексеев, Турыгина, 1974). В омогруппе **CHUCK** имеется три омосегмента: субстантивный (А), глагольный (Б) и междометный (В) (см. рис. 1). Сегменты А и Б объединяют каждый по несколько слов, находящихся между собой в отношениях лексической омонимии, а сегмент В содержит только одно слово; в последнем случае членение омогруппы на сегменты приводит к снятию лексической омонимии.

Для настоящего исследования особый интерес представляют структурные единицы второго типа, объединяющие слова, общие по лексическому значению. Такие единицы мы будем в дальнейшем называть г и п е р л е к с е м а м и (ГЛ). В омогруппе **CHUCK** три гиперлексемы; на рис. 1 они обозначены римскими цифрами (I, II, III).

Понятие гиперлексемы — лексической единицы, охватывающей в парадигматическом плане несколько однокоренных (и тождественных по фонемному составу) [9] слов — необходимо не только для описания явлений омонимии; оно, как это было обстоятельно

[9] Само собой разумеется, что полным это тождество будет только у неизменяемых слов, у изменяемых же речь может идти только о тождестве основных (исходных) форм слов.

аргументировано И. В. Арнольд,[10] имеет важное значение для английской лексикологии вообще. Как показывают подсчеты (см. табл. 9), большой процент английских слов выступает в языке не обособленно, а в составе гиперлексем. Кроме того, в результате происходящих в языке процессов образования новых слов по конверсии практически любое обособленное слово может войти в состав вновь возникшей гиперлексемы. Поэтому в практических целях, в частности при решении задач прикладного характера, удобно считать основным элементом английской лексики не слово, а гиперлексему — двусторонний знак, означающим которого является форма, единая для всех входящих в него слов, а означаемым — общее для них лексическое значение (при этом каждое обособленное слово, не имеющее в языке соотносящихся с ним по конверсии слов, рассматривается как самостоятельная гиперлексема с количеством членов, равным единице). Такой подход может оказаться полезным и в теоретическом плане, поскольку он снимает противоречие между распространенным в зарубежной англистике отношением к слову как к единице, которая может выступать «в функции разных частей речи», и мнением о том, что слово не может принадлежать одновременно к нескольким частям речи (см.: Смирницкий, 1956, с. 72; Воронцова, 1960, с. 27—28; и др.). Попутно заметим, что английские словари, в которых словарная статья соответствует интуитивно понимаемой составителями гиперлексеме (см., например: АРСл; Barnhart, 1974), несомненно, более адекватно отражают структуру словарного состава современного английского языка, чем словари, в которых слова одной гиперлексемы рассматриваются в разных статьях (как, например: БАРС; Webster's Unabridged, 1966; и др.).

Гиперлексему, состоящую из нескольких слов, будем называть с л о ж н о й, поскольку она включает в себя слова разных частей речи, а состоящую из одного слова — п р о с т о й, так как она целиком относится к какой-либо одной части речи. В рассматриваемой омогруппе сложными являются ГЛ I, включающая в себя имя существительное, глагол и междометие с общим семантическим компонентом 'домашняя птица', и ГЛ III, содержащая существительное и глагол с общим значением 'зажимать', а простой — ГЛ II, куда входит существительное со значением 'полено', которое грамматических омонимов не имеет.

Итак, омогруппа делится на несколько омонимичных гиперлексем, каждая из которых в свою очередь состоит из одной или нескольких омонимичных лексем (омонимов). Важно отметить, что если внутри гиперлексемы господствуют отношения чисто грамматической омонимии, то отношения между гиперлексемами носят характер лексической омонимии: каждая гиперлексема

[10] См.: Арнольд, 1966, с. 22, где соответствующее понятие обозначается термином «лексема». Термин «гиперлексема» представляется нам предпочтительным, поскольку он менее отягощен различными значениями и его внутренняя форма точнее передает сущность данного понятия.

находится в лексической оппозиции по отношению к любой другой гиперлексеме. Соответственно этому каждый из членов одной гиперлексемы находится в отношениях лексической омонимии с каждым из членов любой другой гиперлексемы, входящей в состав той же омогруппы. Иначе говоря, слова одной и той же омогруппы, относящиеся к разным гиперлексемам, являются по отношению друг к другу лексическими омонимами — чисто лексическими, если они принадлежат к одной и той же части речи (как, например, слова 1, 4 и 5 или 2 и 6) и лексико-грамматическими, если они относятся к разным частям речи (например, слова 1 и 6, 2 и 5 и т. п.).

Омогруппа может состоять и из одной гиперлексемы. В этом случае гиперлексема обязательно будет сложной, потому что в омогруппе не может быть менее двух членов. В такой минимальной омогруппе лексических омонимов не окажется, она будет состоять только из чисто грамматических омонимов, т. е. сведется к обыкновенному гнезду слов, соотносящихся по конверсии — такому, как, например, LOOK (LOOK *v* 'глядеть', LOOK *n* 'взгляд'). В настоящем исследовании рассматриваются только такие омогруппы, которые обязательно содержат лексические омонимы, т. е. состоят не менее чем из двух гиперлексем.

Подобно гиперлексеме, омогруппа может быть п р о с т о й (если все ее члены принадлежат к одной и той же части речи) или с л о ж н о й (если в ее составе есть слова разных частей речи). В первом случае омогруппа состоит из одного омосегмента, во втором — из нескольких. Омогруппа CHUCK является сложной, а такие омогруппы, как SANDAL (I *n* 'сандалия', II *n* 'сандаловое дерево') или SCRAGGY (I *a* 'тощий', II *a* 'неровный') — простыми.

При изучении омонимических отношений между элементами омогруппы исследуемые единицы удобно сопоставлять попарно. Очевидно, что количество возможных пар элементов равно количеству сочетаний из общего числа элементов в группе по два. Так, если в омогруппе две гиперлексемы, то на уровне гиперлексем в ней возможно лишь одно сопоставление — I : II (одна пара ГЛ), в омогруппе из трех ГЛ — три сопоставления (I : II, I : III, II : III), т. е. три пары ГЛ, в омогруппе из четырех ГЛ — шесть сопоставлений (I : II, I : III, I : IV, II : III, II : IV, III : IV) и т. д. Вообще, количество возможных гиперлексемных пар равно числу сочетаний из n по 2, определяемому по формуле: $A = n(n-1)/2$, где A — число сочетаний и n — количество ГЛ в омогруппе.

Точно так же определяется и количество возможных для данной омогруппы лексемных пар, т. е. пар омонимов. В омогруппе, состоящей из двух омонимичных слов, имеется только одна возможность для бинарного сопоставления элементов (одна омонимическая пара), в омогруппе из трех слов — три возможности (три омонимических пары) и т. д.

В отличие от отношений между членами гиперлексемной пары, которые всегда основаны на оппозиции лексических значений, отношения между членами омонимической пары (омопары) более разнообразны.

Во-первых, члены омопары могут быть противопоставлены друг другу по лексическому значению, по грамматическому значению части речи или по тому и другому одновременно. Омопары, члены которых существенно различаются по лексическому значению, будем называть лексически разнородными, а омопары, члены которых по лексическому значению тождественны — лексически однородными. Омопары, члены которых принадлежат к одной и той же части речи, называются далее простыми, а пары, члены которых относятся к разным частям речи — сложными.[11]

Во-вторых, члены омопары могут входить в одну и ту же гиперлексему (такие омопары мы будем называть внутренними), но могут принадлежать разным гиперлексемам (внешние омопары). Внутренние омопары всегда лексически однородны, так как внутри гиперлексемы нет отношений лексической оппозиции. При этом они обязательно являются сложными, поскольку в гиперлексеме не может быть слов, принадлежащих к одной и той же части речи. Таким образом, члены внутренней омопары всегда находятся между собой в отношениях чисто грамматической омонимии. В отличие от этого внешние омопары всегда лексически разнородны, но в грамматическом плане они могут быть либо простыми, либо сложными. Отсюда следует, что члены внешних омопар всегда являются лексическими омонимами, но в одних случаях это будут чисто лексические омонимы, а в других — лексико-грамматические.

Структура омопары может быть описана формулой, где члены пары обозначены символами частей речи с индексами, указывающими на принадлежность омонима к той или иной гиперлексеме. Например, в омогруппе EGG (I *n* 'яйцо', *v* 'смазывать яичным желтком'; II *v* 'подстрекать') омопары описываются формулами: $n_1 : v_1$, $n_1 : v_2$ и $v_1 : v_2$. Здесь омонимы, находящиеся в отношениях грамматической оппозиции, обозначены разными буквенными символами (n, v), а омонимы, которые находятся в отношениях лексической оппозиции, отмечены разными цифровыми индексами (1, 2). Кроме того, отношения лексической оппозиции эксплицируются знаком двоеточия. Отсюда следует, что формула внешней (лексически разнородной) омопары обязательно содержит знак двоеточия ($n_1 : v_2$, $v_1 : v_2$ и т. п.), а формула внутренней (лексически однородной) пары этого знака не имеет (n_1v_1).

Структурная формула гиперлексем знака двоеточия никогда не содержит, поскольку компоненты гиперлексемы не противостоят друг другу по лексическому значению, например: n_1v_1 (в омогруппе EGG) или $n_1v_1i_1$ (в омогруппе CHUCK). Формула простой гиперлексемы состоит всего лишь из одного символа с индексом, т. е. сводится к формальному обозначению отдельного омонима, например, n_2.

Структура гиперлексемной пары описывается формулой, которая состоит из формул обеих гиперлексем, со знаком двоеточия

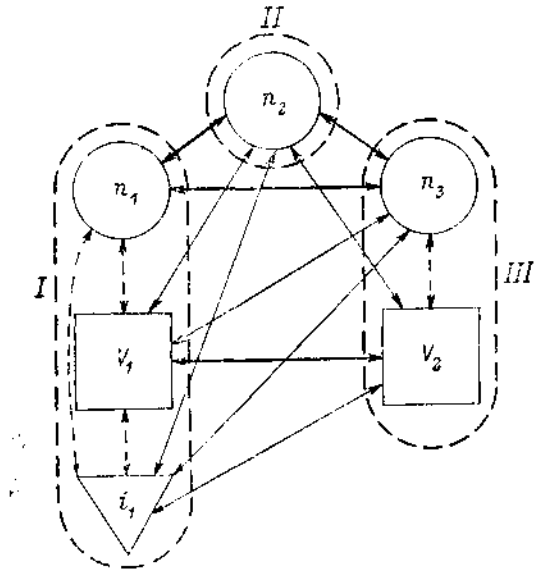


Рис. 2. Структура омогруппы CHUCK на уровне слов.

Условные обозначения: круг — имя существительное, квадрат — глагол, треугольник — междометие. Жирным пунктиром обведены границы гиперлексем (I, II, III). Стрелками показаны отношения между членами омопар (омонимами): пунктирными стрелками соединены члены внутренних, лексически однородных омопар (чисто грамматические омонимы), сплошными жирными — члены внешних грамматически простых омопар (чисто лексические омонимы), сплошными тонкими — члены внешних грамматически сложных омопар (лексико-грамматические омонимы).

между ними, например: $n_1v_1 : v_2$ (в ОГ EGG), $n_1v_1i_1 : n_2 : n_3v_3$ (в ОГ CHUCK) и т. п. Аналогичным образом строится и структурная формула омогруппы: она содержит столько частей (разделяемых знаком двоеточия), сколько в ней имеется гиперлексем. Например, формула омогруппы EGG — $n_1v_1 : v_2$ (т. е. совпадает с формулой гиперлексемной пары), а омогруппы CHUCK — $n_1v_1i_1 : n_2 : n_3v_3$.

На основе описанной классификации рассмотрим омопары, образуемые членами ОГ CHUCK (см. рис. 2).

Слова, входящие в ОГ CHUCK, могут вступать в омонимические отношения пятнадцатью разными способами, т. е. образовывать 15 омопар: $A = \frac{6 \cdot (6-1)}{2} = 15$. Из них 11 омопар являются внеш-

[11] Таким образом, термины «простой» и «сложный» употребляются нами в том же значении, какое вкладывал в них А. И. Смирницкий (см.: Смирницкий, 1956, с. 170).

ними ($n_1 : n_2$, $n_1 : n_3$, $v_1 : v_3$, $v_1 : n_2$, $v_1 : n_3$, $v_1 : v_3$, $i_1 : n_2$, $i_1 : n_3$, $i_1 : v_3$, $n_2 : n_3$ и $n_2 : v_3$) и 4 — внутренними (n_1v_1, n_1i_1, v_1i_1 и n_3v_3).

Все внутренние омопары являются лексически однородными и грамматически сложными, т. е. соответствуют отношениям чисто грамматической омонимии (в указанном выше смысле). Внешние омопары $n_1 : n_2$, $n_1 : n_3$, $n_2 : n_3$ и $v_1 : v_3$ являются лексически разнородными и грамматически простыми, т. е. представляют собой чисто лексические омонимы. Остальные внешние омопары ($n_1 : v_3$, $v_1 : n_2$ и др.) являются лексически разнородными и грамматически сложными и соответствуют, следовательно, отношениям лексико-грамматической омонимии. Всего в омогруппе CHUCK можно выявить 4 пары чисто грамматических, 4 пары чисто лексических и 7 пар лексико-грамматических омонимов.

В английском языке часто встречаются омогруппы, в которых какая-либо часть речи представлена только одним омонимом. Такие омонимы мы будем называть грамматически дифференцированными. Так, в ОГ CASE (I *n* 'случай'; II *n* 'ящик', *v* 'класть в ящик') грамматически дифференцированным омонимом является глагол со значением 'класть в ящик', так как других глаголов в этой омогруппе нет; в ОГ ALIGHT (I *v* 'приземляться'; II *a* 'зажженный') оба омонима дифференцированы грамматически. В ОГ CHUCK к числу дифференцированных омонимов относится междометие CHUCK 'цып!', поскольку других междометий в этой омогруппе не имеется.

Понятие грамматической дифференцированности омонима может оказаться полезным при выяснении того, какую роль в разрешении лексической омонимии может играть грамматический (синтаксический) анализ. В случае грамматической дифференциации установление принадлежности анализируемой омоформы к той или иной части речи одновременно снимает и лексическую неоднозначность.

5. Анализ структуры омонимического ряда. Уровень словоформ

До сих пор исследование характера формально-смысловых отношений между членами омогруппы велось нами на уровне слов, т. е. в качестве минимального компонента омогруппы рассматривалось слово. При этом мы отвлекались от факта изменяемости слов, наличия у многих слов нескольких грамматических форм, различающихся по звучанию. Фактически мы рассматривали основную (словарную) форму слова как представителя всей его парадигмы, пренебрегая формами косвенными. Такой подход позволяет построить лишь весьма упрощенную модель омогруппы. На самом же деле картина формально-смысловых отношений между членами омогруппы значительно сложнее.

Прежде всего, фонетическое (графическое) тождество основных форм омонимичных слов вовсе не обязательно сопровождается

тождеством их косвенных форм (ср., например: LIGHT — lights 'свет' и LIGHT — lighter, lightest 'легкий'). Возможны и такие случаи, когда при тождестве косвенных форм имеет место различие основных (ср., например: SYRINGE — syringes 'шприц' и SYRINX — syringes 'евстахиева труба'). Но даже и тогда, когда слова, входящие в состав омогруппы, обладают тождественными парадигмами, не все их формы находятся в омонимических отношениях друг с другом. Например, в омогруппе ADDER (I *n* 'гадюка', II *n* 'сумматор') парадигмы обоих омонимов тождественны: n_1 *sg* — adder, n_1 *pl* — adders; n_2 *sg* adder, n_2 *pl* — adders; однако из шести возможных пар словоформ омонимичными являются только две — n_1 *sg* : n_2 *sg* и n_1 *pl* : n_2 *pl*; в остальных парах формального тождества нет (adder — adders), а следовательно, нет и омонимии. Полное тождество форм возможно лишь в случае неизменяемых слов, т. е. когда число элементов в каждой из парадигм равно единице.

Ясно, что получить достаточно полное и точное представление о характере формально-смысловых отношений между членами омогруппы можно только в том случае, если сопоставлять между собой не слова в целом, представленные своими словарными формами, а все, в том числе и косвенные, формы слов, т. е. если перейти с уровня слов на уровень словоформ.

При рассмотрении омонимии на уровне словоформ прежде всего встает вопрос о составе парадигм слов изменяемых частей речи.

Нормативная грамматика усматривает в системе английского субстантивного словоизменения две формы числа (единственное и множественное) и две падежные формы — общий и притяжательный падежи (Ganshina, Vasilevskaya, 1958, с. 28). Однако если наличие у существительных двух форм числа ни у кого не вызывает сомнений, то в отношении категории падежа ведется длительная дискуссия. Наиболее обстоятельно аргументированы две точки зрения. Согласно одной из них (см.: Jespersen, 1933, с. 132 сл.; Смирницкий, 1959, с. 124; Жигадло, Иванова, Иофик, 1956, с. 31 сл.), в английском языке имеется два падежа — общий (с нулевым окончанием) и притяжательный (выражаемый формантом -'s); согласно другой — английское существительное вообще не изменяется по падежам, а формант -'s является послелогом (Воронцова, 1960, с. 183).

Основным аргументом против понимания форманта -'s как падежной флексии является то, что -'s может оформлять не только отдельное имя существительное, но и целую атрибутивную конструкцию. Однако если -'s и является послелогом, то это послелог особого рода. Действительно, формант -'s, не будучи флексией, в то же время не обладает самостоятельностью, необходимой для признания его служебным словом, не может существовать отдельно от слова, к которому он примыкает. Именно поэтому он и ощущается часто не как синтаксический служебный элемент (каким он

является на самом деле), а как часть слова, флексия. Этому способствует и то обстоятельство, что в подавляющем большинстве случаев формант -'s следует непосредственно за существительным в атрибутивной функции; употребление его в качестве оформителя атрибутивной группы встречается крайне редко.

В связи с этим представляется целесообразным в программах автоматической переработки текста или при анализе текста в учебных целях упрощенно рассматривать субстантивную словоформу с формантом -'s в качестве особой, притяжательной формы соответствующего существительного. Учитывая факультативность этой формы (передаваемое ею грамматическое значение всегда может быть выражено и без нее, с помощью конструкции с предлогом of), условимся считать формы с -'s «избыточными».

Таким образом, в нормальной парадигме существительного имеется только две формы — *sg* и *pl*; в избыточной же парадигме может быть четыре словоформы — *sg*, *pl*, *sg poss* и *pl poss*. Кроме того, возможны дефектные парадигмы, состоящие только из одной формы числа — либо единственного (например: measles 'корь'), либо множественного (scissors 'ножницы', spectacles 'очки' и др.). Некоторые существительные, имеющие дефектную парадигму, могут в то же время иметь «избыточную» притяжательную форму (например: cattle *pl* — cattle's *pl poss*).

В системе английского адъективного словоизменения обычно различаются три формы, соответствующие трем степеням сравнения — положительной, сравнительной и превосходной. Некоторые зарубежные лингвисты считают степени сравнения не формами одного и того же слова, а разными словами; соответственно и суффиксы степеней сравнения рассматриваются как словообразовательные. А. И. Смирницкий в свое время убедительно показал ошибочность этой точки зрения (Смирницкий, 1959, с. 156—159). При рассмотрении формально-смысловых отношений между омонимами мы будем различать в парадигме прилагательного названные три формы, учитывая при этом, что многие прилагательные — как относительные, так и качественные — обладают дефектной (одночленной) парадигмой, так как либо вообще не изменяются по степеням сравнения в силу особенностей своей семантики (относительные и некоторые качественные, например, main 'главный', middle 'средний' и т. п.), либо образуют степени сравнения аналитически (многосложные прилагательные). В последнем случае дефектность их парадигмы условна, они дефектны только с точки зрения количества синтетических форм.

Те же три формы следует различать и в парадигме наречия, *где так же, как и у прилагательных, часто встречаются слова с дефектной* или условно дефектной парадигмой.

Что касается глагола, то вопрос о количественном и качественном составе его парадигмы весьма сложен, и в рамках настоящей статьи нет возможности остановиться на нем подробно. Поэтому рассмотрим здесь лишь основные результаты проведенного исследования, напомнив, что нас интересуют только синтетические формы.

Следуя за А. И. Смирницким (Смирницкий, 1959, с. 208—361), мы различаем в парадигме глагола 12 синтетических форм (см. табл. 1).

Легко заметить, что в парадигме каждого из приведенных в таблице глаголов много омонимичных словоформ (омоформ). У стандартного глагола CALL 'звать' из двенадцати форм пять (1, 2, 3, 5 и 6) омонимичны друг другу потому, что имеют одинаковое (нулевое) оформление, четыре (7, 8, 9 и 10) — благодаря оформлению одним и тем же суффиксом -ed, две (11 и 12) — потому, что оформлены суффиксом -ing, и только одна (4), оформленная суффиксом -s, не находится в омонимических отношениях ни с какой другой словоформой.

Еще сильнее проявляется омонимия в парадигме глагола HURT 'причинять вред', где одинаковое (нулевое) оформление имеют девять форм (1, 2, 3, 5, 6, 7, 8, 9 и 10). Парадигма глагола BE 'быть', которая является в английском языке максимально дифференцированной, также содержит большое число омонимичных словоформ (1, 2 и 6; 8 и 9; 11 и 12).

Такое явление, когда в парадигме одного и того же слова имеются омонимичные словоформы, будем называть в н у т р и л е к с е м н о й омоформией. Внутрилексемная омоформия всегда является грамматической. Как видно из предыдущего, в английском языке она особенно распространена в системе глагола. У прилагательных и наречий омонимичных словоформ нет; у существительных омонимичные словоформы встречаются сравнительно редко и только в звуковом варианте языка, где совпадают формы *pl*, *sg poss* и *pl poss*.

Сильное развитие омонимии в системе английского глагольного словоизменения связано с тем, что число различных графических комплексов, обслуживающих парадигму глагола, значительно меньше числа словоформ в парадигме. У глагола RING 'звонить' таких комплексов пять — ring, rings, ringing, rang и rung, у глагола CALL — четыре (call, calls, called и calling), у глагола HURT — три (hurt, hurts, hurting), а у глагола BE — восемь (be, am, is, are, was, were, been и being).

Регулярность внутрилексемной омоформии в парадигме глагола приводит к регулярному возникновению м е ж л е к с е м н о й омоформии, т. е. к такому положению, когда наличие омонимических отношений между одной парой словоформ, принадлежащих разным словам, обязательно сопровождается омонимией еще одной или нескольких пар словоформ, относящихся к тем же словам. Так, наличие в омогруппе одной глагольной словоформы со значением инфинитива, омонимичной, скажем, словоформе единственного числа существительного, входящего в эту омогруппу, всегда означает, что в той же омогруппе есть

Таблица 1

Синтетические формы английского глагола

№№ п/п	Значение грамматической формы глагола	Сокращенное обозначение	Примеры парадигм				
			Нестандартные глаголы				Стандартный глагол
			глагол с максимально дифференцированной парадигмой	глагол с дифференциацией форм прошедшего времени и причастия II		глагол с минимально дифференцированной парадигмой	
1	Инфинитив	*inf*	be	ring		hurt	call
2	Повелительное наклонение	*imp*	be	ring		hurt	call
3	1 л. ед. числа — Present Indefinite Active Indicative	*1 p sg*	am	ring		hurt	call
4	3 л. ед. числа — Present Indefinite Active Indicative	*3 p sg*	is	rings		hurts	calls
5	Мн. число — Present Indefinite Active Indicative	*pres pl*	are	ring		hurt	call
6	Сослагательное наклонение I	*sbj I*	be	ring		hurt	call
7	Ед. число — Past Indefinite Active Indicative	*pt sg*	was	rang		hurt	called
8	Мн. число — Past Indefinite Active Indicative	*pt pl*	were	rang		hurt	called
9	Сослагательное наклонение II	*sbj II*	were	rang		hurt	called
10	Причастие II	*pp*	been	rung		hurt	called
11	Причастие I	*pres p*	being	ringing		hurting	calling
12	Герундий	*ger*	being	ringing		hurting	calling

еще по крайней мере четыре глагольных словоформы, находящихся в тех же отношениях с данной субстантивной словоформой.

Подобная регулярность позволяет несколько упростить рассмотрение отношений между словоформами в составе омогруппы. Графически тождественные словоформы одного и того же слова могут быть сведены в «блоки» по две, три и более словоформ, и дальнейшие операции могут производиться уже не с каждой словоформой в отдельности, а с целыми блоками омонимичных между собой словоформ — о м о б л о к а м и. Число омоблоков даже в глагольной парадигме невелико; оно равно числу графических комплексов, обслуживающих парадигму. Так, в парадигме глагола HURT выделяется три омоблока, содержащих от 1 до 9 омоформ каждый (hurt, hurts и hurting), в парадигме глагола RING — пять омоблоков и т. д.

На практике, однако, удобнее иметь дело не с блоком как таковым, а с одной из входящих в него словоформ, произвольно избираемой в качестве представителя всех словоформ блока.[12] Такие словоформы можно называть репрезентативными словоформами (РСФ).

Условимся рассматривать в качестве РСФ стандартного глагола словоформы инфинитива (*inf*), 3 лица ед. числа Present Indefinite Active индикатива (*3 p sg*), ед. числа Past Indefinite Active индикатива (*pt*) и причастия I (*pres p*). В качестве РСФ существительного в графическом варианте языка выступают все четыре формы парадигмы: *sg*, *pl*, *sg poss* и *pl poss*. В парадигмах прилагательного и наречия количество РСФ также совпадает с количеством словоформ. Для неизменяемого слова в качестве РСФ выступает единственная форма этого слова.

Такой подход дает возможность вести анализ структуры омогруппы, не исключая внутрисловную омоформию из рассмотрения, а вынося ее, так сказать, за скобки.

Рассмотрим теперь омогруппу CHUCK,[13] имея в качестве элементарных единиц анализа уже не слова, а словоформы (в необходимых случаях — репрезентативные словоформы).

Если при рассмотрении структуры ОГ CHUCK на уровне слов мы имели перед собой 6 элементарных единиц (слов), которые все были тождественны друг другу по форме, то перейдя на уровень словоформ, мы обнаруживаем уже свыше 30 элементарных единиц — словоформ (6 субстантивных, 24 глагольных и одну междометную), многие из которых не тождественны между собой. Эти элементарные единицы сгруппированы в 15 омоблоков, пред-

[12] Именно так поступают школьные учебники и словари: выделяя так называемые основные формы глагола, они фактически имеют дело с его репрезентативными формами. Однако, в отличие от подхода, принятого в настоящей работе, остальные формы глагола они не рассматривают вообще.

[13] Перечень слов, входящих в состав этой омогруппы, дан на с. 154.

ставленных следующими РСФ (для каждой РСФ в скобках указывается количество представляемых ею словоформ):

chuck n_1 *sg* (1) — chucks n_1 *pl* (1);

chuck v_1 *inf* (5) — chucks v_1 3 *p sg* (1) — chucked v_1 *pt* (4) — chucking v_1 *pres p* (2);

chuck int_1 (1);

chuck n_2 *sg* (1) — chucks n_2 *pl* (1);

chuck n_3 *sg* (1) — chucks n_3 *pl* (1);

chuck v_3 *inf* (5) — chucks v_3 3*p sg* (1) — chucked v_3*pt* (4) — chucking v_3 *pres p* (2).

Как видно из этого перечня, РСФ, входящие в ОГ **CHUCK**, распределяются по нескольким комплектам, различающимся

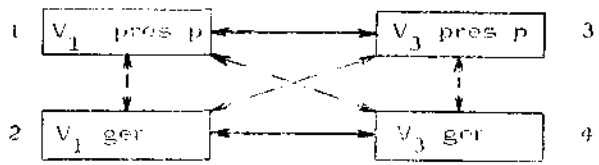


Рис. 3. Структура омокомплекта **chucking**.

Сплошными жирными линиями показаны отношения между грамматически однородными словоформами, пунктирными — между лексически однородными, а штрихпунктирными — между омоформами, образующими омопару, неоднородную как лексически, так и грамматически.

по написанию и звучанию. Внутри каждого такого комплекта господствуют отношения омонимии, так как все РСФ (и соответственно все словоформы), входящие в комплект, тождественны между собой в плане выражения. Поэтому подобные комплекты можно называть о м о к о м п л е к т а м и (ОК). Всего в данной омогруппе четыре ОК:

1) **chuck**, состоящий из 14 словоформ — n_1 *sg* (1), v_1 *inf* (5), int_1 (1), n_2 *sg* (1), n_3 *sg* (1) и v_3 *inf* (5);

2) **chucks**, объединяющий 5 словоформ — n_1 *pl* (1), v_1 3*p sg* (1), n_2 *pl* (1), n_3 *pl* (1) и v_3 3*p sg* (1);

3) **chucked**, содержащий 8 словоформ — v_1 *pt* (4) и v_3 *pt* (4);

4) **chucking**, в который входит 4 словоформы — v_1 *pres p* (2) и v_3 *pres p* (2).

Рассмотрим структуру наиболее простого омокомплекта — **chucking** (рис. 3). В этот омокомплект входит всего две РСФ — v_1 *pres p* и v_3 *pres p*, но поскольку каждая из них соответствует двум словоформам (*pres p* и *ger*), то общее количество словоформ в омокомплекте — четыре. Отсюда число возможных пар омоформ (омопар) — шесть, из которых две внутрилексемные (1—2 и 3—4) и четыре межлексемные (1 : 3, 1 : 4, 2 : 3 и 2 : 4). Межлексемные омопары 1 : 3 и 2 : 4 состоят из словоформ, тождественных по внутреннему грамматическому значению, а пары 1 : 4 и 2 : 3 — из нетождественных. Все омоформы тождественны между собой по внутреннему грамматическому значению.

Омокомплект **chucked** также состоит из двух РСФ (v_1 *pt* и v_3 *pt*), но поскольку каждая из них представляет по 4 словоформы, то общее число словоформ в омокомплекте — 8 (см. рис. 4). Число возможных омопар — 28, из них внутрилексемных 12 (1—2, 1—3, 1—4, 2—3, 2—4, 3—4, 5—6, 5—7, 5—8, 6—7, 6—8 и 7—8), межлексемных с тождеством внутренних грамматических значений — 4 (1 : 5, 2 : 6, 3 : 7 и 4 : 8) и межлексемных с различием этих значений — 12 (1 : 6, 1 : 7, 1 : 8, 2 : 5, 2 : 7, 2 : 8, 3 : 5, 3 : 6, 3 : 8, 4 : 5, 4 : 6 и 4 : 7). Все межлексемные омопары как в этом ОК, так и в предыдущем, лексически разнородны, так как

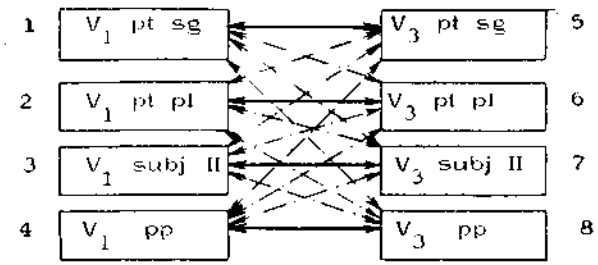


Рис. 4. Структура омокомплекта **chucked**.

Условные обозначения те же, что и на рис. 3, но отношения между лексически однородными омоформами здесь ради простоты не показаны.

члены их относятся к словам, входящим в разные лексемы и не обладающим общностью лексических значений.

Значительно сложнее структура омокомплектов **chucks** и **chuck**. Сложность этих омокомплектов не только в том, что они состоят из большого числа омоформ и содержат больше потенциальных омопар, но прежде всего в том, что в них имеются структурные единицы, промежуточные между омокомплектом и омоформой. Одни из них объединяют омоформы, относящиеся к одной и той же части речи; на уровне слов им соответствуют омосегменты. Другие объединяют омоформы, принадлежащие словам, близким по лексическому значению. Такую единицу (на уровне слов ей соответствует гиперлексема) мы будем называть рабочим термином «формема». В составе омокомплекта формемы всегда омонимичны между собой, т. е. являются омоформемами.

Омоформема представляет собой совокупность омоформ, принадлежащих словам одной гиперлексемы и, следовательно, обладающих общностью лексических значений. Внутри омоформемы все возможные омопары лексически однородны, и наоборот, омопары, члены которых относятся к разным омоформемам, лексически разнородны.

В омокомплекте **chucks** три омоформемы — chucks I, chucks II и chucks III (см. рис. 5). Омоформемы I и II являются сложными, так как содержат каждая более одной омоформы: в омоформему chucks I входят две омоформы — n_1 *pl* и v_1 3*p sg*, близкие друг другу

по лексическому значению, а в омоформему chucks III — также две близкие по лексическому значению омоформы: n_3 *pl* и v_3 *3p sg*. Омоформема chucks II является простой: в нее входит только одна омоформа — n_2 *pl*. Всего в комплекте 5 омоформ, которые могут образовывать 10 различных омопар.

Члены омопары могут принадлежать к одной и той же омоформеме, а могут принадлежать к разным омоформемам. В первом случае мы будем иметь в н у т р и ф о р м е м н у ю пару, во втором — м е ж ф о р м е м н у ю. В омокомплекте **chucks** — две внутриформемные пары (1—2 и 4—5) и 8 межформемных. Среди последних различаются омопары простые, у которых оба члена принадлежат к одной и той же части речи (1 : 3, 1 : 4, 2 : 5 и 3 : 4), и сложные, члены которых относятся к разным частям речи (1 : 5, 2 : 3, 2 : 4 и 3 : 5). Все сложные пары являются в данном случае субстантивно-глагольными.

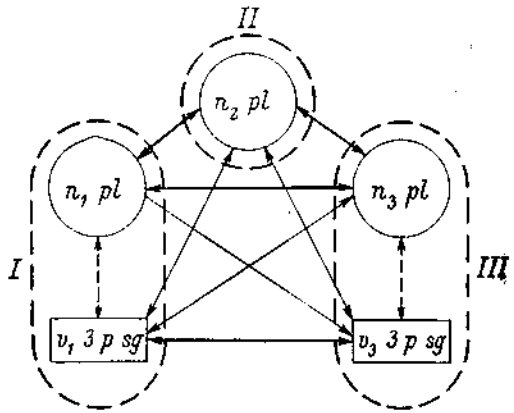


Рис. 5. Структура омокомплекта **chucks**.

Условные обозначения аналогичны обозначениям на рис. 2: субстантивные словоформы изображены в виде кругов, глагольные — в виде прямоугольников. Жирным пунктиром обведены границы формем. Стрелками показаны отношения между членами омопар (омоформами): пунктирными стрелками соединены члены внутриформемных, лексически однородных пар (чисто грамматические омоформы), сплошными жирными — члены внешних грамматически простых и однородных пар (чисто лексические омоформы), сплошными тонкими — члены внешних грамматически сложных пар (лексико-грамматические омоформы).

В омокомплекте **chuck** тоже три омоформемы (см. рис. 6), но общее количество омоформ больше — 14, и отношения между ними сложнее. Здесь представлено уже два типа внутриформемных пар — м е ж б л о ч н ы е, т. е. относящиеся к разным *блокам* (*а следовательно и к разным словам*) *и в н у т р и б л о ч н ы е*. К внутриблочным относятся омопары 2—3, 2—4, 2—5, 2—6, 3—4, 3—5, 3—6, 4—5, 4—6 и 5—6, в каждой из которых оба члена принадлежат глаголу chuck v_1, и омопары 10—11, 10—12, 10—13, 10—14, 11—12, 11—13, 11—14, 12—13, 12—14 и 13—14, оба члена которых принадлежат глаголу chuck v_3 (всего 20 пар). К межблочным относятся омопары 1—2, 1—3, 1—4, 1—5, 1—6, 1—7, 2—7, 3—7, 4—7, 5—7, 6—7 (в составе омоформемы chuck I) и 9—10, 9—11, 9—12, 9—13, 9—14 (в составе омоформемы chuck III).

Как и в омокомплекте **chucks**, межформемные омопары образуют также два типа — пары простые и пары сложные. При этом простые омопары имеются двух видов: грамматически однородные, члены которых т о ж д е с т в е н н ы по внутренним грамматическим значениям (числа, лица, времени и т. п.) — 1 : 8, 1 : 9, 8 : 9, 2 : 10, 3 : 11, 4 : 12, 5 : 13, 6 : 14, и грамматически разнородные, члены которых р а з л и ч а ю т с я по внутренним грамматическим значениям — 2 : 11, 2 : 12, 2 : 13, 2 : 14, 3 : 10, 3 : 12, 3 : 13, 3 : 14, 4 : 10, 4 : 11, 4 : 13, 4 : 14, 5 : 10, 5 : 11, 5 : 12, 5 : 14, 6 : 10, 6 : 11, 6 : 12 и 6 : 13. Среди сложных омопар в данном омокомплекте помимо пар субстантивно-глагольных (1 : 10, 1 : 11, 1 : 12, 1 : 13, 1 : 14, 2 : 8, 2 : 9, 3 : 8, 3 : 9, 4 : 8, 4 : 9, 5 : 8, 5 : 9, 6 : 8, 6 : 9, 8 : 10, 8 : 11, 8 : 12, 8 : 13 и 8 : 14) имеются также субстантивно-междометные (7 : 8 и 7 : 9) и глагольно-междометные (7 : 10, 7 : 11, 7 : 12, 7 : 13 и 7 : 14) омопары.

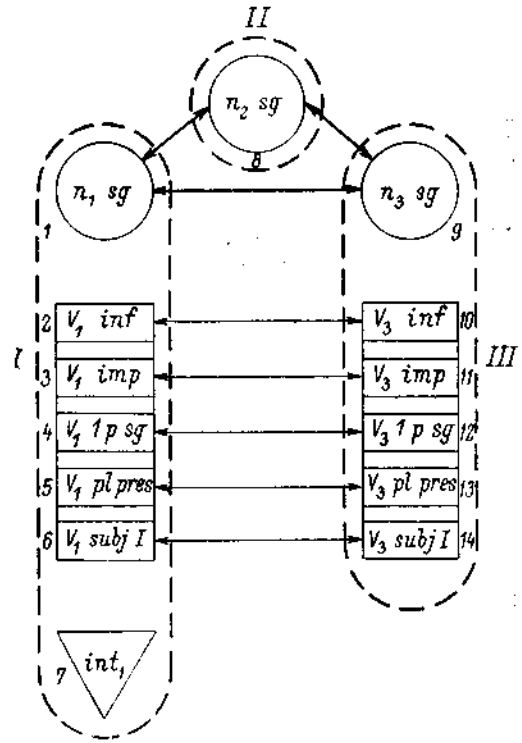


Рис. 6. Структура омокомплекта **chuck**.

Треугольником представлена междометная словоформа; остальные обозначения те же, что и на рис. 5. В целях упрощения показаны отношения только между грамматически однородными омоформами.

Всего в омокомплекте **chuck**, таким образом, может быть образована 91 омопара нескольких различных типов. Количество блоков, словоформ и потенциальных омопар в каждом из омокомплектов группы **CHUCK**, а также во всей омогруппе в целом, представлено в таблице 2, а распределение их по типам — в таблице 3.

Таблица 2

Количество блоков, словоформ и потенциальных омопар в омогруппе **CHUCK**

Омокомплект	Количество блоков	Количество *словоформ*	Количество потенциальных омопар
chuck	6	14	91
chucks	5	5	10
chucked	2	8	28
chucking	2	4	6
Во всей ОГ	15	31	135

6. Разграничение смежных омогрупп. Уточнение определения омогруппы

Омогруппа **CHUCK** является в некотором смысле идеальной омогруппой: в ней нет ни одной словоформы, которая не находилась бы в омонимических отношениях с какой-либо другой словоформой этой же омогруппы, и в то же время ни одна из словоформ *этой группы не находится в омонимических отношениях со словоформами* какой-либо иной омогруппы. Иначе говоря, все словоформы ОГ **CHUCK** являются омоформами, при этом омонимические отношения, в которых состоят эти словоформы, не выходят за пределы данной омогруппы.

Такое идеальное положение наблюдается далеко не всегда. Во многих омогруппах имеются «лишние», не состоящие в омонимических отношениях между собой, члены. Так, в ОГ DATE (n_1 'дата', 'срок', v_1 'датировать; назначать свидание', n_2 'финик') глагольные словоформы dating и dated не омонимичны никаким другим словоформам:

DATE n_1 *sg* — dates *pl*;

DATE v_2 *inf* — dates *3p sg* — dating *pres p* — dated *pt*.

Часто встречаются и такие омогруппы, у которых отдельные члены вступают в омонимические отношения со словами из других омогрупп. Это хорошо видно на примере группы LIE. В эту омогруппу входят слова LIE n_1 'ложь', LIE v_1 'лгать', LIE n_2 'логово' и LIE v_2 'лежать' (см. рис. 7).

Таблица 3

Распределение потенциальных пар омоформ в омогруппе **CHUCK** по типам

Омокомплект	Внутриформенные омопары (ĈD)			Межформенные омопары (Ĉ)						Итого
	Всего	В том числе:		Всего	В том числе:					
		внутрисловные (ĈDd)	межсловные (ĈDd)		простые (ĈD)		сложные (ĈD)			
					грамматически тождественные (ĈDd)	грамматически различные (ĈDd)	$n : v$	$n : int$	$v : int$	
chuck	36 2 12 2	20 — 12 2	16 2 — —	55 8 16 4	8 4 4 2	20 — 12 2	20 4 — —	2 — — —	5 — — —	**91** **10** **28** **6**
Во всей ОГ	52	34	18	83	18	34	24	2	5	135

Условные обозначения: Ĉ — сходство (**C** — существенное различие) лексических значений; D — тождество (D — различие) внешних грамматических значений; d — тождество (d — различие) внутренних грамматических значений.

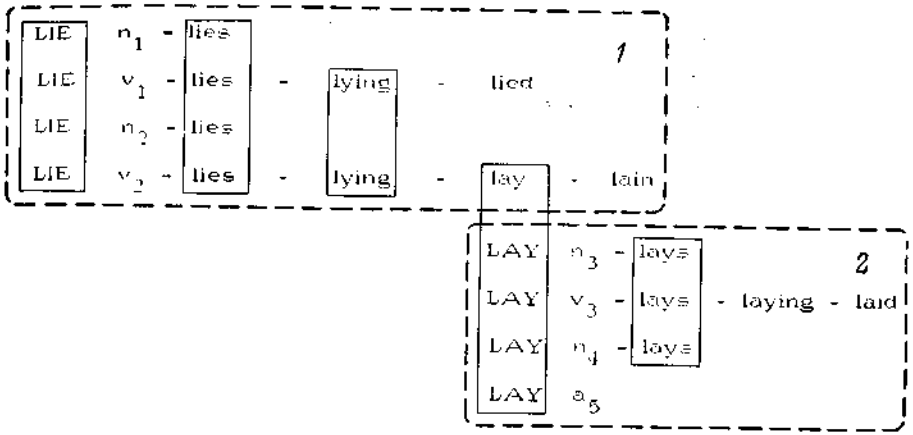


Рис. 7. Омонимические отношения между формами слов, входящих в омогруппу LIE (1) и омогруппу LAY (2).

Предположительные границы омогрупп показаны пунктирной линией. Прописными буквами даются основные (словарные) формы слова, а строчными — косвенные (несловарные).

Словоформы, имеющие написание LIE, lies и lying, группируются в три омокомплекта, а словоформы lied (v_1 *pt*), lay (v_2 *pt*) и lain (v_2 *pp*) являются в н е к о м п л е к т н ы м и, т. е. не находятся в омонимических отношениях ни с какой другой словоформой данной омогруппы. Однако словоформа lay оказывается омонимичной нескольким словоформам, входящим в ОГ LAY (n_1 'положение', 'расположение', v_1 'класть'; n_2 'короткая песенка'; *a* 'светский').

Возникает вопрос: как провести разграничительную линию между ОГ LIE и ОГ LAY? С одной стороны, ясно что глагол LIE v_2 должен быть включен в ОГ LIE, поскольку некоторые из его форм (LIE, lies, lying) совпадают со многими формами слов LIE n_1, LIE v_1 и LIE n_2, являющимися членами этой омогруппы. С другой стороны, представляется, что этот же глагол должен войти и в омогруппу LAY, так как одна из его форм (lay) совпадает с формами слов LAY n_3, LAY v_3, LAY n_4 и LAY a_5, образующими омогруппу LAY, а согласно определению, сформулированному в начале статьи, для вхождения в омогруппу достаточно совпадения слов хотя бы в одной из форм. Таким образом, получается, что глагол LIE v_2 входит одновременно в две омогруппы — омогруппу LIE и омогруппу LAY, т. е. что омогруппы эти являются пересекающимися множествами (см. рис. 8). Правда, можно было бы попытаться решить этот вопрос иначе, предположив, что омогруппы LIE и LAY образуют одну большую омогруппу, членом которой и является глагол LIE v_2. Но от такого решения приходится отказаться, потому что если объединить эти омогруппы в одну, то членами этой единой омогруппы окажутся такие слова,

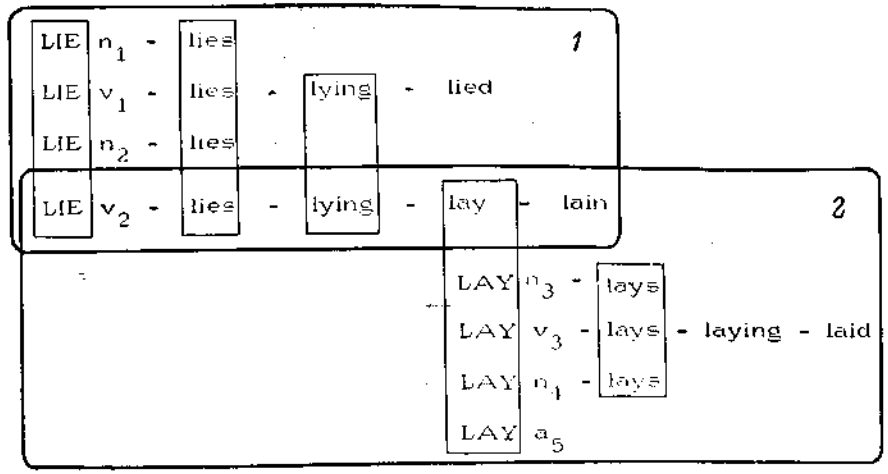


Рис. 8. Взаимопересечение омогрупп LIE (1) и LAY (2).

Границы омогрупп обозначены жирной линией.

которые не совпадают ни в одной из своих форм (например, LIE n_1 и LAY n_3, LIE n_1 и LAY v_3, LIE v_1 и LAY n_4 и т. д.), что противоречит нашему определению омогруппы.

К сожалению, не избавляет нас от противоречия и первое решение — считать, что глагол LIE v_2 входит и в омогруппу 1 (LIE), и в омогруппу 2 (LAY). Действительно, если слово LIE v_2 принадлежит одновременно омогруппе 1 и омогруппе 2, то тогда и все словоформы этого слова должны принадлежать обеим этим омогруппам, т. е. словоформа LIE v_2 *inf* должна входить как в омогруппу 1, так и в омогруппу 2, словоформы lies v_2 *3p sg*, lying v_2 *pres p*, lay v_2 *pt* и lain v_2 *pp* также должны входить в обе эти ОГ, и т. д. Принадлежность словоформы lay v_2 *pt* обеим этим омогруппам не вызывает сомнения, поскольку графический комплекс lay имеется и в омогруппе 1, где он соответствует внекомплектной словоформе прошедшего времени глагола LIE v_2, и в омогруппе 2, где он соответствует четырем омонимичным исходным формам слов — членов группы. Но если допустить, что и остальные четыре словоформы глагола LIE v_2 (LIE, lies, lying и lain) принадлежат обеим омогруппам, то придется признать, что все омонимичные им словоформы других слов (LIE n_1 *sg*, lies n_1 *pl*, LIE v_1 *inf* и т. д.) тоже принадлежат не только омогруппе 1, но и омогруппе 2. В результате получится, что омогруппа 2 целиком поглотит омогруппу 1, а главное — что в одной и той же омогруппе окажутся слова, не имеющие ни одной совпадающей формы (например, LIE n_1 и LAY n_3, LIE v_1 и LAY a_5 и т. д.), т. е. не являющиеся омонимами. Следовательно, мы придем к тому же противоречию, что и раньше.

Таким образом, ни объединять омогруппы **LIE** и **LAY**, ни включать глагол LIE v_2 в состав обеих омогрупп нельзя. Как же тогда провести границу между омогруппами?

Очевидно граница должна проходить таким образом, чтобы словоформа lay этого глагола, являющегося членом омогруппы 1, входила бы одновременно и в омогруппу 2, а исходные формы слов LAY n_3, LAY v_3, LAY n_4 и LAY a_5, являющихся членами омогруппы 2, входили бы одновременно и в омогруппу 1. Омогруппы, таким образом, будут действительно пересекающимися, но в качестве общей части у них будет выступать не слово

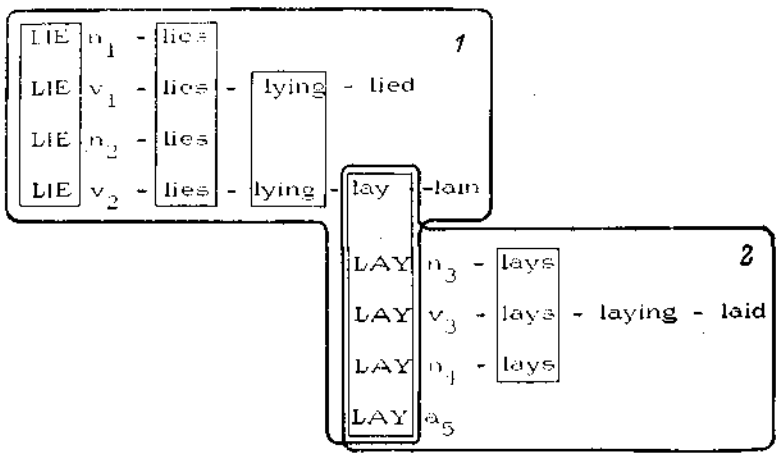


Рис. 9. Уточнение границ между омогруппами **LIE** (1) и **LAY** (2).

LIE v_2, как мы предполагали вначале, и не одна лишь словоформа lay v_2 *pt*, а весь омокомплект lay/**LAY** (см. рис. 9).

Итак, мы получили четкое разграничение пересекающихся омогрупп. Однако теперь возникло новое противоречие: некоторые омонимы оказались разнесенными в разные омогруппы. Так, слово LIE v_2 попало в омогруппу **LIE**, а совпадающее с ним в одной из форм слово LAY n_3 — в омогруппу **LAY**. В омогруппу **LAY** попали и другие частичные омонимы глагола LIE v_2. По определению же омонимы, в том числе и частичные, должны входить в одну и ту же омогруппу.

Поскольку описанный подход к разграничению омогрупп дает в каждом конкретном случае вполне удовлетворительное, однозначное решение этой задачи, отказываться от него, очевидно, не следует; необходимо, по-видимому, пересмотреть наше определение омогруппы, которое было сформулировано применительно к уровню слов и не могло еще учитывать данных, полученных при изучении омогруппы на уровне словоформ.

Рассмотрим подробнее, какие словоформы содержатся в омогруппах **LIE** и **LAY**. Как видно из рис. 9, в состав каждой из

этих двух ОГ входят омонимичные между собой основные (словарные) формы слов, далее — косвенные формы, соответствующие этим основным, и, наконец, что очень важно, формы, не принадлежащие парадигмам этих слов, но находящиеся в отношениях омонимии с каким-либо членом (или членами) этих парадигм. Так, в омогруппу 1 входят основные формы слов LIE n_1, LIE v_1, LIE n_2 и LIE v_2, образующие омокомплект, затем соответствующие им косвенные формы, объединенные в омокомплекты **lies** и **lying**, косвенные формы, остающиеся вне комплектов — lied v_1 *pt* и lain v_2 *pp*, и, наконец, косвенная форма lay v_2 *pt* и омонимичные ей четыре словоформы омогруппы **LAY**, образующие все вместе единый омокомплект. В свою очередь, в омогруппу 2 входят омонимичные друг другу основные формы слов LAY n_3, LAY v_3, LAY n_4 и LAY a_5, соответствующие им косвенные словоформы, образующие омокомплект **lays**, внекомплектные косвенные словоформы laying v_3 *pres p* и laid v_3 *pt*, а также словоформа lay v_2 *pt*, входящая в один омокомплект с уже перечисленными основными формами слов данной омогруппы.

Мы видим, таким образом, что при установлении состава омогруппы важную роль играет различие основных и косвенных форм слов: если слова омонимичны друг другу в основных формах, они входят в одну и ту же ОГ, если же слова омонимичны между собой только в косвенных формах, то такие слова попадают в разные ОГ. В последнем случае наличие омонимических отношений находит свое выражение в том, что омонимичные словоформы включаются в один и тот же омокомплект, который является для этих разных омогрупп общим.

Теперь нам ясно, что омонимы могут быть членами одной и той же омогруппы, а могут входить и в разные омогруппы. Полные омонимы всегда являются членами одной и той же омогруппы, а частичные — лишь в том случае, если у них совпадают основные формы; в случае же совпадения только косвенных форм или основной формы одного слова с косвенной формой другого (как, например, у слов LIE v_2 и LAY n_3) частичные омонимы попадают в разные — но пересекающиеся — омогруппы.

Сказанное позволяет нам более точно сформулировать определение омогруппы. Омогруппа — это сложное образование, включающее в себя, во-первых, все слова языка, тождественные между собой фонетически и/или графически в своих основных (а иногда и косвенных) формах, но существенно различающиеся по своим лексическим и/или грамматическим значениям, и, во-вторых, все те формы других слов языка, которые тождественны каким-либо формам данных слов. Таким образом, слова, совпадающие в основных формах (словарные омонимы), входят в состав омогруппы полностью, всеми своими парадигмами, тогда как из парадигм несловарных омонимов в ее состав включаются лишь отдельные словоформы.

Основные формы словарных омонимов образуют основной,

или ядерный, омокомплект омогруппы. Подобно тому, как основная форма слова является как бы представителем всего слова (Ахманова, 1957, с. 108), всей его парадигмы, так и ядерный омокомплект представляет омогруппу в целом, служит ее «лицом». Вокруг ядерного омокомплекта группируются остальные, периферийные элементы ОГ — омокомплекты, состоящие из косвенных форм слов, и внекомплектные словоформы.

Ядерный омокомплект является для омогруппы обязательным, без него омогруппа не существует; периферийные же элементы являются факультативными и могут отсутствовать.

7. Типы словарности омонимического ряда

Таким образом, омонимические ряды — как омогруппы, так и омокомплекты — могут различаться по своей «словарности», т. е. по наличию или отсутствию в них словарных форм и по роли этих форм в данном ряду. Этот факт имеет большое значение для разработки методики поиска омонимических рядов по словарю, а также для их последующего количественного описания. Поэтому мы остановимся на типах словарности подробнее.

Прежде всего, омокомплекты могут быть словарными, т. е. состоять только из словарных форм слов (см. выше ОК **chuck**) и несловарными — образуемыми косвенными формами слов (ср. ОК **chucks**, **chucking**, **chucked**). Возможны и смешанные омокомплекты, в которых словарные формы сочетаются с несловарными. Среди смешанных ОК следует различать комплекты, базирующиеся на омонимии словарных форм, комплекты, базирующиеся на несловарных формах, и симметричные ОК. В первом случае речь идет об омокомплектах, образованных двумя или несколькими словарными формами, к которым примыкает одна (редко более) несловарная форма (например, $CLOVE_1$ *n sg* 'гвоздика' — $CLOVE_2$ *n sg* 'долька чеснока' — $clove_3$ *pt, pp* от CLEAVE *v* 'раскалывать(ся)'), во втором — о комплектах, образованных двумя или более несловарными формами, к которым примыкает одна словарная (например, $boots_1$ *pl* от BOOT *n* 'ботинок' — $boots_2$ *3 p sg* от BOOT *v* 'помогать' — $BOOTS_3$ *n sg* 'коридорный'), а в третьем — о комплектах, в которых одна словарная форма сочетается с одной несловарной (например, DIVERS *a* 'разный' — divers *pl* от DIVER *n* 'ныряльщик').

Соответственно этому омогруппы также подразделяются на словарные, несловарные и смешанные. Словарная омогруппа это такая, в которой в качестве ядерного ОК выступает словарный омокомплект. В такой омогруппе, независимо от наличия или отсутствия в ее составе несловарных омокомплектов, все омонимы обязательно представлены своими словарными формами (как, например, в ОГ CHUCK), благодаря чему словарные омогруппы легко обнаруживаются в обычном словаре. В смешанной омогруппе ядерным является смешанный омокомплект. Поиск таких омогрупп в обычном словаре затруднителен, поскольку некоторые из их членов представлены лишь косвенными, несловарными формами. В несловарной омогруппе ядерного омокомплекта фактически нет. Несловарная ОГ состоит из единственного омокомплекта, образуемого несловарными формами таких слов, которые в словарных формах не омонимичны, например, $bases_1$ *pl* от BASE *n* 'база' и $bases_2$ *pl* от BASIS *n* 'базис'. Подобный омокомплект выступает в языке автономно и — за отсутствием сопряженного с ним словарного омокомплекта — должен рассматриваться как самостоятельная омогруппа, а слова, формы которых в нем представлены — как полноправные (хотя и частичные) омонимы. Выявление несловарных омогрупп по обычному словарю невозможно, необходима особая методика, о которой будет сказано ниже.

Омогруппы могут подразделяться на единичные, состоящие из одного омокомплекта (как, например, ОГ **ALIGHT** I *a* 'зажженный', II *v* 'спешиваться'), и множественные, состоящие из нескольких сопряженных ОК (как, например, рассматривавшиеся выше ОГ **CHUCK**). Во множественной группе омокомплекты связаны отношениями производности: ядерный (словарный или смешанный с преобладанием словарной омонимии) омокомплект является производящим, а периферийные (несловарные и смешанные с преобладанием несловарной омонимии) — производными. В единичной группе, где есть только ядерный омокомплект и нет периферийных, ядерный ОК является непроизводящим (например, ОК **alight**), а в тех редких случаях, когда в омогруппе нет ядерного омокомплекта и она состоит из одного несловарного ОК (ср. ОГ **BASES**) — автономным.

Следует, наконец, упомянуть о явлении вторичной словарности, которое заключается в том, что в несловарный омокомплект входит лексикализовавшаяся косвенная форма слова. Чаще всего это существительное или прилагательное, образовавшееся от глагольной формы с суффиксом -ing или -ed. Так, например, в омокомплект **closing**, помимо причастных форм от трех глаголов — $CLOSE_1$ 'закрывать', $CLOSE_2$ 'заканчивать и $CLOSE_3$ 'сближаться', входят отглагольное существительное $CLOSING_1$ 'закрытие' и прилагательное $CLOSING_2$ 'заключительный'; в омокомплект **baited**, помимо форм прошедшего времени и причастия II от глаголов $BAIT_1$ 'приманивать' и $BAIT_2$ 'травить', входит прилагательное BAITED 'затравленный'. К «вторично-словарным» следует отнести и такие формы, как BOUND *a* 'связанный' (от причастия II глагола BIND 'связывать') или BETTER *n* 'вышестоящее лицо' и BETTER *v* 'улучшать', восходящие к сравнительной степени **better** прилагательного GOOD и т. п.

Несловарные ОК с вторично-словарными формами следовало бы, может быть, из чисто теоретических соображений относить к числу словарных, однако по своей структуре они отличаются от обычных словарных ОК и сходны с несловарными, вследствие чего их количественный учет и лексикографическая обработка значительно упрощаются, если считать их не словарными комплектами, а особой разновидностью комплектов несловарных. Поэтому в дальнейшем мы будем относить их к числу «несловарных омокомплектов с вторичной словарностью», а входящие в них словарные формы учитывать особо как «вторично-словарные формы».

8. Основные результаты изучения структуры омонимического ряда

При рассмотрении уровня слов выявляются омонимические ряды лексем — омогруппы. В качестве основных частей омогруппы выделяются, с одной стороны, омосегменты — единицы, противостоящие друг другу по грамматическому значению части речи, а с другой — гиперлексемы, единицы, находящиеся между собой в отношениях лексической оппозиции. Внутри омосегмента минимальные члены омогруппы — омонимы — тождественны по грамматическому значению и противопоставлены по лексическому, а внутри гиперлексемы — сходны по лексическому значению и различаются по грамматическому.

При рассмотрении уровня словоформ выявляются омонимические ряды словоформ — омокомплекты. Они обнаруживаются как в составе омогрупп, так и при сопоставлении парадигм слов, не омонимичных в своих словарных формах. Омогруппа предстает на уровне словоформ как совокупность сопряженных омокомплектов, с одной стороны, и внекомплектных словоформ (т. е. таких, которые не находятся в отношениях омонимии ни с какими иными словоформами) — с другой. Омокомплект выступает здесь как основная структурная единица омонимики, обладающая собственной внутренней структурой.

Структура омокомплекта во многом аналогична структуре омогруппы. Прежде всего, омокомплект, подобно омогруппе, членится на части, противостоящие друг другу по внешнему грамматическому значению (аналоги омосегментов). На членении этого рода мы подробно останавливаться не будем; отметим только, что внутри каждой такой части минимальные члены омокомплекта — омоформы — тождественны между собой по категориальному грамматическому значению и различаются по лексическому. Далее, в составе омокомплекта выделяются части, противопоставленные по лексическому значению — формемы (аналоги гиперлексем), которые, в свою очередь, делятся на омоблоки (аналоги омонимов), а каждый омоблок содержит одну или

несколько омоформ. Внутри формемы омоблоки, подобно омонимам в гиперлексеме, сходны между собой по лексическому значению и различаются по грамматическому значению части речи; а внутри омоблока все омоформы тождественны друг другу по лексическому значению и по значению части речи и различаются только грамматическими значениями числа, лица, времени и т. п.

Полезно теперь сопоставить выявленные нами основные структурные единицы омонимики языка — омогруппу и омокомплект.

Прежде всего, омогруппа — образование весьма разнородное: в качестве его элементов выступают единицы разных языковых уровней — как целые слова, так и отдельные словоформы. Омокомплект же представляет собой совокупность однородных единиц — словоформ.

Далее, в омогруппе могут быть такие элементы, которые не тождественны друг другу по форме, т. е. не находятся в омонимических отношениях; в омокомплекте все элементы обязательно тождественны между собой. Иначе говоря, члены омогруппы могут быть полными омонимами, а могут находиться в отношениях частичной или неравнообъемной омонимии, тогда как по отношению к элементам омокомплекта вопрос о полноте или частичности омонимии вообще не встает: здесь омонимия либо есть, либо ее нет совсем.

Омонимические отношения между членами омогруппы могут выходить за пределы данной омогруппы: омонимами (частичными) могут быть и члены разных омогрупп; омонимические же отношения между элементами омокомплекта полностью замыкаются в нем самом, поскольку он включает в себя в с е словоформы языка, омонимичные между собой.

И, наконец, омогруппы являются пересекающимися множествами, омокомплекты же никогда не пересекаются.

Все это показывает, что в основу полного исчисления омонимов должны быть положены не омогруппы, а омокомплекты.[14] Иными словами, точное описание омонимики языка может быть получено только при условии обращения к уровню словоформ.

Однако описание омонимики в терминах омокомплектов не может рассматриваться как окончательный этап исследования омонимии. В связи с тем, что отношения между омокомплектами и омогруппами сложны и запутаны (в одних случаях омокомплект является частью омогруппы, в других — соответствует целой омогруппе, в третьих — входит одновременно в две разные омогруппы, а иногда выступает совершенно самостоятельно,

[14] Именно неразработанность этого вопроса и отсутствие представления об омокомплекте как об основной структурной единице омонимики на уровне словоформ и явились причиной того, что до сих пор не было получено адекватное количественное описание омонимики языка.

автономно от каких бы то ни было омогрупп), описание омонимики языка в терминах омокомплектов не может дать четкого представления о характере омонимических отношений между словами, входящими в омокомплекты своими отдельными формами. Для всестороннего понимания этих отношений необходимо описание омонимики в терминах омогрупп, т. е. обратный переход с уровня словоформ на уровень слов.

Вкратце процедура этого перехода сводится к выявлению в массиве омокомплектов языка словарных омокомплектов и поиску сопряженных с ними (т. е. содержащих формы тех же слов) несловарных омокомплектов. Каждый словарный омокомплект кладется в основу формируемой омогруппы, образует ее ядро, а сопряженные с ним несловарные омокомплекты (вместе с внекомплектными словоформами) — ее периферию. Смешанный омокомплект включается в состав сразу двух омогрупп, выступая в одной из них в качестве ядерного, а в другой — в качестве периферийного комплекта (как, например, омокомплект LAY/lay на рис. 9). В результате мы получаем два параллельных списка омонимических рядов языка — список омокомплектов и список омогрупп, причем все элементы одного списка будут четко сопоставлены с соответствующими элементами другого.

II. Квантитативные характеристики омонимических рядов

1. История вопроса

Вопрос о количественной характеристике явлений омонимии уже давно интересует исследователей (см., например: Philipon-la-Madelaine, 1802; Bridges, 1919). Много ли омонимов в языке, какова их доля в общем словарном составе, как различаются по числу омонимов разные языки, как распределяются омонимы по частям речи, какова зависимость между омонимией и средней длиной слова или его употребительностью — вот некоторые из вопросов, рассматривавшихся исследователями на материале английского, русского и других языков (Liebich, 1898; Jespersen, 1928, с. 356 сл.; Zipf, 1935, с. 30 сл.; Костюченко, 1954; Урбутис, 1956; Конецкая, 1961; Шайкевич, 1962; Алексеев, Турыгина, 1974, с. 1—6; Тышлер, 1975; и др.).

Однако результаты, полученные разными авторами, во многом противоречивы. Так, резко различаются цифры общего количества омонимов в английском языке: 1800 слов по данным У. Скита (Skeat, 1909), 2947 — по подсчетам А. Я. Шайкевича (Шайкевич, 1962, с. 275), свыше 4800 — по Ю. П. Костюченко (Костюченко, 1954), 2500 или 5096 — по подсчетам И. С. Тышлера, сделанным в разное время (Тышлер, 1960; 1975); расходятся данные о доле гомогенных омонимов среди общего числа омонимов (24% по подсчетам Ю. П. Костюченко, 7% по данным В. П. Конецкой, 1.5% по мнению И. С. Тышлера (см.: Тышлер, 1963, с. 43)), о соотношении числа односложных и многосложных омонимов (по данным А. Я. Шайкевича односложных омонимов в 2.3 раза больше, чем многосложных (Шайкевич, 1962, с. 279), по мнению И. С. Тышлера (Тышлер, 1960, с. 80) — в 9 раз) и мн. др.

Расхождения эти объясняются не только различиями в критериях выделения омонимов. В не меньшей степени они связаны с отсутствием четкого представления о структуре омонимического ряда и видах омонимических единиц и проистекающим отсюда разнобоем в используемых единицах счета. В одних случаях подсчитываются графические омонимы, в других — фонетические, одни авторы учитывают только полные омонимы, другие включают в подсчет и частичные и т. д. Положение особенно ухудшается тем, что часто в одной и той же работе смешиваются единицы разных уровней и содержательные выводы делаются на основе сопоставления разных, несопоставимых между собой, данных. Так, в работе Ю. П. Костюченко при определении числа омографов учитывается и лексическая, и грамматическая омонимия, а при подсчете числа фонетико-графических *омонимов и омофонов — только лексическая*. В этой же работе, а также в упоминавшихся уже работах И. С. Тышлера подсчет общего числа омонимов ведется по списку, который состоит из единиц разных уровней — лексем и словоформ. Яркий пример подобного смешения уровней можно встретить в работе И. С. Тышлера (Тышлер, 1975, с. 73—75), где делается попытка определить процент омонимов среди наиболее частотных слов английского языка. По подсчетам И. С. Тышлера, использовавшего частотный словарь Торндайка-Лорджа (Thorndike, Lorge, 1944), «у первой, самой высокой по употребительности тысячи слов, свыше 250 — омонимы (более 25%)». Однако единицей счета в списке Тышлера являются «омонимы», т. е. слова, тогда как в словаре Торндайка-Лорджа в качестве основных единиц выступают не слова, а гиперлексемы, а иногда и целые омогруппы. Поэтому фактически первой тысяче вокабул этого словаря соответствует по крайней мере две тысячи разных слов, и, следовательно, процент омонимов снизится вдвое. Если же учесть, что среди «омонимов» в списке И. С. Тышлера приводятся не только слова, но и словоформы, то станет ясно, что сопоставлять его с данными словаря Торндайка-Лорджа вообще нельзя.

Из сказанного следует, что квантитативное описание омонимики может быть корректным только в том случае, если оно строится на основе четкого понимания структуры омонимического ряда и при последовательном разграничении единиц разных уровней. Именно поэтому в настоящей работе количественному изучению омонимики современного английского языка предшествовало детальное исследование структуры омонимического ряда.

2. *Методика исследования*

Количественное изучение омонимики проводилось нами на материале лексических омонимов: рассматривались только такие омонимические ряды, которые состоят из чисто лексических или лексико-грамматических омонимов; чисто грамматическая же («моделированная») омонимия охватывалась лишь постольку, поскольку она связана с омонимией лексической, сопровождает ее. Так, омонимы WORK *v* 'работать' — WORK *n* 'работа' в исследование не включались, тогда как подобные же омонимы BLOW *v* 'думать' — BLOW *n* 'дуновение' рассматривались, потому что они оба находятся в отношениях лексической омонимии со словами $BLOW_1$ *n* 'цвет' и $BLOW_2$ *n* 'удар', а также $BLOW_2$ *v* 'ударять'.

Выделение лексических омонимов и выявление омонимических рядов (омогрупп) производилось на основе «Краткого Оксфордского словаря» (COD, 1964) с привлечением в необходимых случаях близких ему по объему, но более четко подающих грамматическую и лексико-семантическую структуру слова малого словаря Уэбстера (Webster, 1971 — ниже WCD) и «Англо-русского словаря» под редакцией В. К. Мюллера (АРСл, 1971). Из словаря Уэбстера черпались главным образом малоупотребительные производные слова с высокопродуктивными суффиксами -er, -ly и -ness, а также отглагольные существительные и прилагательные на -ing и -ed, которые обычно опускаются Оксфордским словарем ради экономии места. АРСл оказался особенно полезен тем, что в нем очень четко разграничены лексические и грамматические омонимы, которые зачастую в COD и в WCD подаются без разграничения.

В связи с тем, что исследование ориентировано на британский вариант английского языка, формы, характерные для американского употребления, принимались во внимание только в том случае, если они приводятся в COD.

Из словарей выбирались только графические омонимы (как чисто графические так и фонетико-графические); выявление чисто фонетических омонимов (омофонов) в задачи исследования, как указывалось выше, не входило.

Разграничение омонимии и полисемии осуществлялось в соответствии с данными всех трех словарей. В большинстве случаев эти данные совпадали; в случае же расхождения предпочтение отдавалось тому (или тем) из словарей, где дается более дифференцированная картина. Это позволило в некоторой степени противостоять обычному для словарей английского языка консерватизму в оценке случаев расхождения значений слова.

С помощью указанной методики в словарях было обнаружено свыше 1500 омонимических групп, содержащих около 5000 омонимов. Для подробного количественного исследования были взяты омонимы, содержащиеся в отрезке алфавита от A до E

включительно, что составляет приблизительно 30% общего объема словарного состава языка. В этом отрезке было выявлено 443 омогруппы, т. е. все словарные омогруппы и смешанные омогруппы с преобладанием словарной омонимии.

Для выявления несловарной омонимии (несловарных омокомплектов и несловарных форм смешанных омогрупп) требуется в идеале составить полный алфавитный список словоформ (точнее РСФ) языка. Эта работа чрезвычайно трудоемка: объем такого списка составил бы (по всему алфавиту) не менее 130 тысяч единиц. Поэтому поиск несловарных и смешанных омонимов был проведен по упрощенной методике. Был составлен список словоформ не для всех слов исследуемого отрезка, а только для выявленных омонимов. В результате была получена картотека омоформ (омонимичных РСФ) объемом свыше трех тысяч единиц. Сопоставление этих РСФ между собой позволило выявить почти всю несловарную омонимию — производные омокомплекты, а также несловарные формы смешанных омокомплектов, а сопоставление их с корпусом словарей — смешанную омонимию. Кроме того, некоторые смешанные омонимы были выявлены благодаря принятому в английской лексикографии правилу включать нерегулярные косвенные формы слов в общий алфавит словаря. Это позволило пополнить список смешанных омонимов такими словами, как DUG *n* 'сосок (животного)' — dug *pt, pp* от DIG *v* 'копать', BETTER *n* 'держащий пари' — better *comp* от GOOD *a* 'хороший', CREW *n* 'судовая команда' — crew *pt* от CREW *v* 'кукарекать'. Некоторое количество смешанных ОК, а также несловарных ОК, не связанных ни с какими словарными, удалось выявить, выбирая из словарей пары слов, близких по написанию в исходной форме и тождественных в форме множественного числа — AXE и AXIS, CINERARIA и CINERARIUM, BELLOW и BELLOWS, DO и DOE и т. п. (ср. соответствующие формы множественного числа: axes, cineraria, bellows, does).

Всего с помощью описанных приемов было выявлено 33 смешанных и 4 автономных ОК, что позволило довести список омогрупп до 480 единиц; общее же число омокомплектов составило 1096. Эти две совокупности, охватывающие приблизительно одну треть всей омонимики английского языка (как на уровне слов, так и на уровне словоформ), выделяемой на базе словаря объемом в 60—80 тысяч слов, и явились основным материалом для квантитативного описания омонимических рядов.

Были получены следующие квантитативные характеристики:

1) распределение омонимических рядов по типам словарности (табл. 4 и 5);

2) распределение омогрупп по количеству омокомплектов, гиперлексем и омонимов, а также по количеству сегментов (табл. 6, 7, 8 и 13);

3) распределение омокомплектов по количеству формем и омоблоков (табл. 10 и 11);

4) распределение гиперлексем по количеству омонимов и формем по количеству омоблоков (табл. 9 и 12);

5) распределение омогрупп по структурным типам (табл. 14—17);

6) распределение межлексемных омопар по структурным типам (табл. 18 и 19);

7) распределение гиперлексем по лексико-грамматическим разрядам и омонимов по частям речи (табл. 20, 21, 23 и 24);

8) соотношение количества простых и сложных омогрупп, *гиперлексем и межлексемных омопар (табл. 22);*

9) доля грамматически дифференцированных омонимов и омоформ (табл. 25);

10) распределение омокомплектов по длине словоформы и средняя длина омоформы (табл. 26);

11) омонимическая насыщенность словаря (табл. 27);

12) частотное распределение омонимических рядов (табл. 28 и 29);

13) оценка доли омокомплектов в общем количестве словоформ языка и ее зависимость от частоты (табл. 30);

14) зависимость объема омокомплекта и формемы от частоты ОК (табл. 31);

15) распространенность грамматической дифференциации омонимов в разных частотных зонах (табл. 32);

16) зависимость между длиной омонима (омоформы) и частотой омонимического ряда (табл. 33—34).

Таблица 4

Распределение омогрупп по типам словарности

№ типа	Типы ОГ	Количество ОГ данного типа	
		в абсолютных цифрах	в %% к общему числу ОГ
1	Словарные ОГ	436	90.8
2	Несловарные ОГ (автономные омокомплекты)	4	0.8
3	Смешанные ОГ,	40	8.4
	в том числе:		
	а) основанные на омонимии словарных форм	(7)	(1.5)
	б) основанные на омонимии несловарных форм	(12)	(2.5)
	в) симметричные	(21)	(4.4)
	Всего	480	100.0

3. Количественная характеристика типов словарности

Изучение распределения омогрупп по типам словарности (см. табл. 4) показало, что подавляющее большинство ОГ в английском языке (более 90%) является словарными, т. е. такими, все члены которых представлены в обычном словаре. На долю смешанных ОГ приходится 8.4%, а на долю несловарных — менее 1% (омогруппы **axes**, **does**, **butted** и **butting**). Такое преобладание словарной омонимии, по-видимому, типично для языков с бедной морфологией и большой регулярностью словоизменения; наоборот, в таком языке, как русский, можно ожидать существен-

Таблица 5

Распределение омокомплектов по типам словарности

№ типа	Типы ОК	Количество ОК данного типа			
		неосложненных	с вторичной словарностью	всего	
				в абсолютных цифрах	в %% к общему числу ОК
1	Словарные ОК:				
	а) производящие	259	0	259	23.6
	б) непроизводящие	87	0	87	8.0
	в) с производящими и непроизводящими членами	90	0	90	8.2
	Всего словарных ОК	436	0	436	39.8
2	Несловарные ОК:				
	а) производные	527	87	614	56.0
	б) непроизводные (автономные)	4	0	4	0.4
	в) с производными и непроизводными членами	2	0	2	0.2
	Всего несловарных ОК	533	87	620	56.6
3	Смешанные ОК:				
	а) с преобладанием словарных форм, в том числе:				
	производящие	3	2	5	0.4
	непроизводящие	2	0	2	0.2
	б) с преобладанием несловарных форм, в том числе:				
	производные	9	3	12	1.1
	непроизводные	0	0	0	0.0
	в) симметричные	18	3	21	1.9
	Всего смешанных ОК	32	8	40	3.6
	Итого: в абсолютных цифрах	1001	95	1096	100.0
	Итого: в %% к общему числу ОК	91.3	8.7	100.0	—

ного преобладания несловарной и смешанной омонимии (ср. *мой* — местоимение и императив глагола, *жало* — существительное и глагол в прошедшем времени, *три*, *трем* — числительное и глагол, *горю* — существительное и глагол и т.п.).

Рассмотрение типов словарности на уровне омокомплектов (см. табл. 5) показало, что несловарные ОК составляют большинство (около 60%, если считать также ОК типа 3б и 3в). Чтобы полнее осветить картину взаимоотношения между словарными и несловарными ОК, было подсчитано количество «производящих» и «непроизводящих» словарных ОК (т. е. связанных и не связанных отношениями формообразования с какими-либо несловарными ОК) и «производных» и «непроизводных» несловарных ОК (т. е. связанных и не связанных отношениями формообразования с какими-либо словарными ОК). Общее число производящих ОК (типы 1а, 1в и часть типа 3а) составило 354, а производных (типа 2а, 2в и 3б) — 628. Таким образом, среди омокомплектов, связанных отношениями формообразования, на один производящий (словарный или смешанный) ОК приходится в среднем около двух производных. Из этих цифр, кроме того, следует, что отношениями формообразования охвачено подавляющее большинство ОК — 982 из 1096, т. е. около 90%. Среди словарных ОК — а они играют ведущую роль в формировании омогрупп — полностью или частично производящие ОК (типы 1а и 1в) составляют большинство (около 80%). Среди несловарных ОК подавляющее большинство (99%) являются производными.

Более полная картина охвата омокомплектов деривационными отношениями выявляется на основании данных таблицы 6.

Таблица 6

Распределение омогрупп по количеству омокомплектов в группе

Объем ОГ в омокомплектах	Количество ОГ данного объема		Всего ОК	
	в абсолютных цифрах	в %%	в абсолютных цифрах	в %%
1	124	25.8	124	11.2
2	217	45.2	434	39.2
3	9	1.9	27	2.5
4	129	26.9	516	46.6
5	0	0.0	0	0.0
6	1	0.2	6	0.5
Всего	480	100.0	1107	100.0
В среднем на одну ОГ			2.31	

Примечание. Суммарное число ОК в данной таблице больше действительного числа омокомплектов (1096), потому что некоторые ОК входят одновременно в две омогруппы.

Из общего числа омогрупп лишь 124 (около 26%) являются одиночными, т. е. состоят из членов, омонимичных только в одной форме; все остальные содержат по два и более сопряженных омокомплекта каждая. Наличие двух сопряженных ОК типично для субстантивной омогруппы, трех — для адъективной, а четырех — для глагольной. Максимальное число сопряженных ОК составило 6 (субстативно-глагольная омогруппа с омонимией притяжательных форм), а среднее число ОК в омогруппе — 2.31.

4. Объем омонимического ряда

Объем омогрупп измерялся в гиперлексемах (см. табл. 7) и в омонимах (см. табл. 8). Выяснилось, что в большинстве случаев (свыше 90%) объем омогруппы составляет 2 или 3 ГЛ; реже встречаются омогруппы объемом в 4—6 ГЛ, и всего один раз встретилась омогруппа (BUCK), состоящая из 9 ГЛ. Средний объем

Таблица 7

Распределение омогрупп по количеству гиперлексем в группе

Объем ОГ в гиперлексемах	Количество ОГ данного объема		Всего гиперлексем	
	в абсолютных цифрах	в %%	в абсолютных цифрах	в %%
2	350	72.9	700	61.2
3	96	20.0	288	25.2
4	22	4.6	88	7.7
5	8	1.7	40	3.5
6	3	0.6	18	1.6
9	1	0.2	9	0.8
Всего	480	100.0	1143	100.0
В среднем на одну ОГ			2.37	

омогруппы в английском языке 2.37 ГЛ, что, по-видимому, значительно меньше, чем в некоторых других языках, например, в японском (ср.: Pei, 1966, с. 95). Максимальное число омонимов в омогруппе составило 13, но большинство омогрупп содержит 2, 3 или 4 омонима. Среднее число омонимов в ОГ — 3.11. Сравнение таблицы 8 с данными И. С. Тышлера, полученными на материале фонетических (чисто фонетических и фонетико-графических) омонимов (Тышлер, 1975, с. 368), показывает, что среди графических омонимов чаще встречаются группы, содержащие по 2 омонима (40.8% вместо 32% у Тышлера) и реже — группы,

Таблица 8

Распределение омогрупп по количеству омонимов в группе

Объем ОГ в омонимах	Количество ОГ данного объема		Всего омонимов	
	в абсолютных цифрах	в %%	в абсолютных цифрах	в %%
2	196	40.8	392	26.1
3	148	30.9	444	29.5
4	72	15.0	288	19.2
5	34	7.1	170	11.3
6	16	3.3	96	6.4
7	7	1.5	49	3.3
8	5	1.0	40	2.6
10	1	0.2	10	0.7
13	1	0.2	13	0.9
Всего	480	100.0	1502	100.0

состоящие из 3 и 4 омонимов (30.9% и 15.0% вместо 35% и 19% у Тышлера).

К этим данным примыкают сведения о распределении гиперлексем по числу омонимов (см. табл. 9). Максимальное число

Таблица 9

Распределение гиперлексем по количеству омонимов в гиперлексеме

Объем ГЛ в омонимах	Количество ГЛ данного объема		Количество омонимов	
	в абсолютных цифрах	в %%	в абсолютных цифрах	в %%
1	823	72.0	823	54.8
2	290	25.4	580	38.6
3	24	2.1	72	4.8
4	4	0.3	16	1.1
5	1	0.1	5	0.3
6	1	0.1	6	0.4
Всего ГЛ	1143	100.0	1502	100.0
В среднем в одной ГЛ			1.31	

омонимов в одной ГЛ составило 6, но подавляющее большинство ГЛ (97%) состоит из одного (простые ГЛ) или двух омонимов. Средний объем гиперлексем оказался равным 1.31 омонима. Эта

величина характеризует степень участия лексических омонимов в отношениях чисто грамматической омонимии, а также, по-видимому, и степень развития чисто грамматической омонимии в английском языке вообще, если рассматривать совокупность лексических омонимов как случайную выборку из словаря (вы-

Таблица 10

Распределение омокомплектов по количеству формем в комплекте

Объем ОК в формемах	Количество ОК данного объема		Всего формем	
	в абсолютных цифрах	в %%	в абсолютных цифрах	в %%
2	847	77.2	1694	67.0
3	196	17.9	588	23.3
4	32	2.9	128	5.1
5	15	1.4	75	3.0
6	4	0.4	24	0.9
9	2	0.2	18	0.7
Всего	1096	100.0	2527	100.0
В среднем на один ОК			2.31	

Таблица 11

Распределение омокомплектов по количеству омоблоков в комплекте

Объем ОК в омоблоках	Количество ОК данного объема		Всего омоблоков	
	в абсолютных цифрах	в %%	в абсолютных цифрах	в %%
2	503	45.9	1006	31.2
3	348	31.7	1044	32.4
4	134	12.4	536	16.6
5	63	5.7	315	9.8
6	30	2.7	180	5.6
7	9	0.8	63	1.9
8	6	0.5	48	1.5
10	1	0.1	10	0.3
11	1	0.1	11	0.3
13	1	0.1	13	0.4
Всего	1096	100.0	3226	100.0
В среднем на один ОК			2.94	

борочная проверка по отрезку A—E в АРСл дала довольно близкую цифру — 1.23).

Распределение омокомплектов по количеству формем (см. табл. 10) и омоблоков (см. табл. 11) оказалось, как и следовало *ожидать, сходным с распределением ОГ по количеству ГЛ и* омонимов, только соответствующие кривые идут несколько круче, т. е. слегка завышены на оморядах малого объема (2 и 3 структурные единицы) и занижены на рядах большого объема. Более низким оказалось и среднее количество структурных единиц, приходящихся на один омонимический ряд. Все это свидетельствует о том, что для несловарных форм (которых среди ОК *большинство) характерен меньший объем омонимического* ряда, чем для словарных. О том же говорят и данные о количестве омоблоков в формеме, приведенные в таблице 12 (построенной аналогично табл. 9, но описывающей уровень словоформ).

Было получено также распределение омогрупп по количеству сегментов, т. е. грамматически однородных частей омогруппы (см. табл. 13). В изученном материале имеются омогруппы, состоящие из 5 или 6 сегментов, но большинство омогрупп (около 90%) *состоит из одного (простые ОГ) или двух сегментов.*

Таблица 12

Распределение формем по количеству омоблоков в формеме

Объем формемы в омоблоках	Количество формем данного объема						Количество омоблоков		
	в словарных и смешанных ОК		в несловарных ОК		всего				
	в абсолютных цифрах	в %/%	в абсолютных цифрах	в %/%	в абсолютных цифрах	в %/%	в словарных ОК	в несловарных ОК	всего
1	823	72.0	1051	75.9	1874	74.16	823	1051	1874
2	290	25.4	326	23.6	616	24.37	580	652	1232
3	24	2.1	7	0.5	31	1.23	72	21	93
4	4	0.3	—	—	4	0.16	16	—	16
5	1	0.1	—	—	1	0.04	5	—	5
6	1	0.1	—	—	1	0.04	6	—	6
Всего формем	1143	100.0	1384	100.0	2527	100.00	1502	1724	3226
В среднем в одной формеме							*1.31*	*1.25*	*1.28*

Примечание. Распределение формем в словарных и смешанных омокомплектах по количеству омоблоков тождественно распределению гиперлексем в омогруппах по количеству омонимов (ср. табл. 9), поскольку каждая формема в таком ОК («словарная формема») выступает как представитель соответствующей гиперлексемы на уровне словоформ.

Таблица 13

Распределение омогрупп по количеству сегментов в группе

Объем ОГ в сегментах	Количество ОГ данного объема		Количество сегментов	
	в абсолютных цифрах	в %/%	в абсолютных цифрах	в %/%
1	167	34.8	167	19.3
2	255	53.1	510	59.0
3	47	9.8	141	16.3
4	9	1.9	36	4.2
5	1	0.2	5	0.6
6	1	0.2	5	0.6
Всего ОГ	480	100.0	864	100.0
В среднем в одной ОГ			1.80	

5. Распределение омонимических единиц различных уровней по структурным типам

Изучение структуры омогруппы выявило большое разнообразие структурных типов омогрупп в английском языке. Число различных типов ОГ в изученном материале составило 89 (см. табл. 14). Характерно, что большинство из них (56%) отличаются низкой встречаемостью (1), т. е. являются уникальными, нетипичными для языка. С другой стороны, очень небольшое число типов характеризуется высокой повторяемостью: уже первый по частоте структурный тип $(n : n)$[15] охватывает 25% общего числа ОГ, а первые 5 типов — свыше 50%. Подавляющее большинство структурных типов являются субстантивными или глагольными; без участия существительных или глаголов образованы всего 5 типов из 89.

Еще большим разнообразием отличаются структурные типы ОГ, выделенные с учетом внутренних грамматических значений, т. е. учитывающие грамматические значения имеющихся в ряде ОГ несловарных омоформ (см. табл. 15). Таких типов оказалось 115, причем бо́льшая часть из них (73) являются уникальными, а следовательно, не характерными для языка. Из общего числа

[15] В настоящем параграфе используются следующие условные обозначения: a — прилагательное, a_r — прилагательное в сравнительной степени, b — наречие, b_r — наречие в сравнительной степени, c — союз, i — междометие, n — существительное, n_s — существительное во множ. числе, o — местоимение, t — артикль, v — глагол, v_{ed} — глагол в прошедшем времени, v_{-a} — причастие I, v_s — глагол в 3-м лице ед. числа настоящего времени.

Таблица 14

Распределение омогрупп по структурным типам

№ типа	Тип ОГ	Количество ОГ	№ типа	Тип ОГ	Количество ОГ
1	2	3	1	2	3
1	*n : n*	121	46	*n : nv : nv : v*	1
2	*n : nv*	69	47	*nv : nv : v*	1
3	*n : v*	25	48	*nv : v : v*	1
4	*nv : nv*	19	49	*n : n : n : n : n : nva*	1
5	*nv : v*	18	50	*n : n : n : nv : na*	1
6	*n : n : n*	17	51	*n : n : nva*	1
7	*n : a*	17	52	*n : n : v : a*	1
8	*n : n : nv*	16	53	*n : nv : na*	1
9	*v : v*	14	54	*n : nv : va*	1
10	*n : nv : nv*	13	55	*n : nva : a*	1
11	*n : na*	10	56	*nv : nv : va : a*	1
12	*n : nv : v*	9	57	*nv : nv : va*	1
13	*n : n : v*	7	58	*nv : va*	1
14	*n : nva*	7	59	*na : v*	1
15	*a : a*	7	60	*n : n : n : n : nv : nvab*	1
16	*n : n : n : nv*	4	61	*n : vab : b*	1
17	*n : n : nv : nv*	4	62	*nv : nva : vab*	1
18	*n : n : a*	4	63	*n : nvb*	1
19	*nv : nv : nv*	4	64	*nv : nv : nvb*	1
20	*n : va*	3	65	*nv : nvb*	1
21	*nv : na*	3	66	*n : n : nv : ai*	1
22	*nv : a*	3	67	*n : nv : nvbi*	1
23	*na : a*	3	68	*n : nv : nv : nvi*	1
24	*n : ab*	3	69	*n : nv : nvi*	1
25	*n : i*	3	70	*n : nv : v : i*	1
26	*n : n : n : n*	2	71	*nv : i*	1
27	*n : n : n : n : nv*	2	72	*n : ni*	1
28	*n : n : n : v*	2	73	*n : i*	1
29	*n : n : nv : v*	2	74	*n : nv : nvabp*	1
30	*n : v : v*	2	75	*n : n : nvp*	1
31	*n : nv : a*	2	76	*n : v : p*	1
32	*nv : nva*	2	77	*na : bp*	1
33	*n : n : na*	2	78	*n : p*	1
34	*n : nvab*	2	79	*n : nvbco*	1
35	*n : nvi*	2	80	*n : bco*	1
36	*v : v : v*	2	81	*n : p : p : t*	1
37	*v : a*	2	82	*v : v : a*	1
38	*a : a : a*	2	83	*v : va : a*	1
39	*n : n : n : n : n*	1	84	*v : b*	1
40	*n : n : n : n : nv : nv : : nv : nv : v*	1	85	*vb : bp*	1
			86	*v : vp*	1
41	*n : n : n : n : nv : v*	1	87	*a : b*	1
42	*n : n : n : nv : nv*	1	88	*b : b : b*	1
43	*n : n : nv : nv : nv*	1	89	*c : t*	1
44	*n : n : nv : v : v*	1			
45	*n : n : v : v : v*	1		Всего	480

Примечание. При перечислении типов с равной встречаемостью соблюдается следующая иерархия частей речи (соответствующая степени их участия в омонимических отношениях): существительное, глагол, прилагательное, наречие, междометие, предлог, союз, местоимение, артикль.

Таблица 15

Распределение омогрупп по структурным типам (с учетом внутренних грамматических значений)

№ типа	Тип ОГ	Количество ОГ	№ типа	Тип ОГ	Количество ОГ
1	2	3	1	2	3
1	*n : n*	117	48	*n : n : nv : v : v*	1
2	*n : nv*	65	49	*n : n : v : v : v*	1
3	*nv : nv*	19	50	*n : nv : nv : v*	1
4	*n : v*	18	51	*n : v : v*	1
5	*n : n : n*	17	52	*nv : nv : v*	1
6	*n : a*	17	53	*nv : v : v*	1
7	*nv : v*	16	54	*n : n : n : n : n : nva*	1
8	*n : n : nv*	14	55	*n : n : n : nv : na*	1
9	*n : nv : nv*	12	56	*n : n : nva*	1
10	*v : v*	12	57	*n : n : v : a*	1
11	*n : na*	10	58	*n : nv : na*	1
12	*n : nv : v*	8	59	*n : nva : a*	1
13	*a : a*	7	60	*nv : nv : va*	1
14	*n : nva*	6	61	*nv : va*	1
15	*n : n : v*	5	62	*na : v*	1
16	*n : n : a*	4	63	*n : n : n : n : nv : nvab*	1
17	*n : n : n : nv*	3	64	*n : nvab*	1
18	*n : n : nv : nv*	3	65	*n : vab : b*	1
19	*nv : nv : nv*	3	66	*na : nva : vab*	1
20	*nv : na*	3	67	*n : nvb*	1
21	*na : a*	3	68	*nv : nv : nvb*	1
22	*n : ab*	3	69	*nv : nvb*	1
23	*n : i*	3	70	*n : n : nv : ai*	1
24	$n_s : v_s$	3	71	*n : nv : nvbi*	1
25	*n : n : n : n*	2	72	*n : nv : nv : nvi*	1
26	*n : n : n : n : nv*	2	73	*n : nv : nvi*	1
27	*n : n : n : v*	2	74	*n : nv : v : i*	1
28	*n : nv : a*	2	75	*nv : i*	1
29	*n : va*	2	76	*n : ni*	1
30	*nv : nva*	2	77	*n : i*	1
31	*nv : a*	2	78	*n : nv : nvabp*	1
32	*n : n : na*	2	79	*n : n : nvp*	1
33	*n : nvi*	2	80	*na : bp*	1
34	*v : v : v*	2	81	*n : p*	1
35	*v : a*	2	82	*n : nvbpco*	1
36	*a : a : a*	2	83	*n : bco*	1
37	$n : n_s$	2	84	*n : p : p : t*	1
38	$n_s : n_s$	2	85	*v : b*	1
39	$n : n_s v_s$	2	86	*vb : bp*	1
40	$n_s : n_s : n_s v_s$	2	87	*a : b*	1
41	$n : v_{pt}$	2	88	*b : b : b*	1
42	$nv : v_{pt}$	2	89	*c : t*	1
43	*n : n : n : n : nv : nv : : nv : nv : v*	1	90	$n_s : n_s : n_s : n_s : n_s$	1
			91	$n : n_s v_s : v_s$	1
44	*n : n : n : n : nv : v*	1	92	$n_s : n_s : n_s : n_s v_s$	1
45	*n : n : n : nv : nv*	1	93	$n : n_s : n_s v_s : n_s v_s$	1
46	*n : n : nv : nv : nv*	1	94	$n_s : n_s v_s : n_s v_s$	1
47	*n : n : nv : v*	1	95	$n_s : n_s v_s$	1

Таблица 15 (*продолжение*)

1	2	3	1	2	3
96	$n : n : v_{ng}$	1	106	$n : nv_{pt}a$	1
97	$n : nv_{ng}$	1	107	$nv : nv : v_{pt}a : a$	1
98	$n : v_{ng} : v_{ng}$	1	108	$nv : a_r$	1
99	$n : v_{ng}$	1	109	$n : nva_rb_r$	1
100	$n : n : nv : v_{pt}$	1	110	$n : v_{ng} : p$	1
101	$n : n : v_{pt}$	1	111	$v_{pt} : v_{pt}$	1
102	$nv : nv : nv_{pt}$	1	112	$v_{pt} : v_{pt}a : a$	1
103	$n : v_{pt}$	1	113	$v : v_{pt} : a$	1
104	$n : nv_{ng} : v_{ng}a$	1	114	$v_{ng} : v_{ng}$	1
105	$n : v_{ng}$	1	115	$v_{ng} : v_{ng}p$	1
				Всего	480

Примечание. Здесь и далее в таблицах 17, 19, 21 и 24 при перечислении типов с равной встречаемостью символы несловарных омоформ даются в следующем порядке: n_s (множ. число существительного), v_s (3-е лицо ед. числа Present Indefinite), v_{ng} (причастие I), v_{pt} (Past Indefinite), a_r и b_r (сравнительные степени прилагательного и наречия).

типов образованы с участием несловарных омоформ лишь 33, а состоят из одних несловарных омоформ только 8.

Для целей автоматического анализа текста представляет интерес распределение омогрупп по «редуцированным» структурным типам. Формула редуцированного типа содержит только символы разных частей речи для слов, образующих омогруппу, не указывая на количество этих слов и на характер отношений между ними. Эта формула может быть получена из формулы полного структурного типа путем теоретико-множественного сложения входящих в нее символов. Так, тип $n : n$ редуцируется до n, тип $n : nv$ —

Таблица 16

Распределение омогрупп по редуцированным структурным типам

№ типа	Тип ОГ	Количество ОГ	№ типа	Тип ОГ	Количество ОГ	№ типа	Тип ОГ	Количество ОГ	№ типа	Тип ОГ	Количество ОГ
1	*nv*	205	8	*nvab*	5	15	*nvbi*	1	22	*vb*	1
2	*n*	141	9	*ni*	5	16	*nvabp*	1	23	*vbp*	1
3	*na*	36	10	*va*	4	17	*nabp*	1	24	*vp*	1
4	*nva*	31	11	*nvb*	3	18	*np*	1	25	*ab*	1
5	*v*	16	12	*nab*	3	19	*nvbpco*	1	26	*b*	1
6	*a*	9	13	*nvp*	2	20	*nbco*	1	27	*ct*	1
7	*nvi*	6	14	*nvai*	1	21	*npt*	1		Всего	480

до *nv*, тип $nv : nva : vab$ — до *nvab* и т. д. Формула редуцированного структурного типа ОГ характеризует наиболее сжатым образом ее амбивалентность — способность ее членов выступать в качестве слов, относящихся к разным частям речи.

Изученные ОГ распределились по 27 редуцированным структурным типам (см. табл. 16), причем и в этом случае очень многие из них оказались уникальными, а небольшое их число показало чрезвычайно высокую встречаемость. Так, к первым двум типам (*nv* и *n*) относится свыше 70% общего числа ОГ. Как видно из таблицы, здесь также преобладают типы субстантивные и глагольные.

Распределение омогрупп по редуцированным структурным типам, проведенное с учетом внутренних грамматических значений, выявило 44 разных типа (см. табл. 17), также характери-

Таблица 17

Распределение омогрупп по редуцированным структурным типам с учетом внутренних грамматических значений

№ типа	Тип ОГ	Количество ОГ	№ типа	Тип ОГ	Количество ОГ	№ типа	Тип ОГ	Количество ОГ	№ типа	Тип ОГ	Количество ОГ
1	*nv*	181	13	nvv_{pt}	4	25	*nvp*	1	37	$nvv_{pt}a$	1
2	*n*	136	14	*nvb*	3	26	*nabp*	1	38	$nv_{pt}a$	1
3	*na*	36	15	*nab*	3	27	*np*	1	39	nva_r	1
4	*nva*	26	16	n_s	3	28	*nvbpco*	1	40	nva_rb_r	1
5	*v*	14	17	nv_{pt}	3	29	*nbco*	1	41	$nv_{ng}p$	1
6	*a*	9	18	*va*	2	30	*npt*	1	42	vv_{pt}	1
7	n_sv_s	8	19	nn_s	2	31	*vb*	1	43	v_{ng}	1
8	*nvi*	6	20	$nv_{ng}a$	2	32	*vbp*	1	44	$v_{ng}p$	1
9	*ni*	5	21	$v_{pt}a$	2	33	*ab*	1			
10	*nvab*	4	22	*nvai*	1	34	*b*	1		Всего	480
11	nn_sv_s	4	23	*nvbi*	1	35	*ct*	1			
12	nv_{ng}	4	24	*nvabp*	1	36	nv_{pt}	1			

зующихся наличием большого числа уникальных типов, высокой встречаемостью наиболее частотных типов и преобладанием субстантивно-глагольной омонимии.

Аналогичным образом были распределены по структурным типам все входящие в состав изученных омогрупп внешние (межлексемные) омопары, т. е. пары лексических и лексико-грамматических омонимов (см. табл. 18 и 19). Так же, как и в распределении омогрупп, здесь отмечается резкое преобладание субстантивных и глагольных омонимических отношений над всеми прочими; так, доля типов $n : n$, $n : v$ и $v : v$ в общем числе типов омопар составляет около 85%. Однако разнообразие структурных типов

Таблица 18

Распределение внешних (межлексемных) пар омонимов по структурным типам

№ типа	Тип омопары	Количество пар	№ типа	Тип омопары	Количество пар	№ типа	Тип омопары	Количество пар	№ типа	Тип омопары	Количество пар
1	*n : n*	667	7	*n : i*	20	13	*b : b*	4	19	*n : t*	1
2	*n : v*	549	8	*a : a*	17	14	*a : b*	3	20	*a : p*	1
3	*v : v*	158	9	*n : p*	10	15	*n : c*	2	21	*p : p*	1
4	*n : a*	98	10	*v : b*	9	16	*n : o*	2	22	*p : c*	1
5	*v : a*	42	11	*v : i*	7	17	*b : p*	2	23	*p : o*	1
6	*n : b*	24	12	*v : p*	5	18	*p : t*	2	24	*c : t*	1
										Всего	1627

Таблица 19

Распределение внешних (межлексемных) пар омонимов по структурным типам (с учетом внутренних грамматических значений)

№ типа	Тип омопары	Количество пар	№ типа	Тип омопары	Количество пар	№ типа	Тип омопары	Количество пар	№ типа	Тип омопары	Количество пар
1	$n : n$	629	12	$n : p$	10	23	$v : p$	3	34	$a : p$	1
2	$n : v$	501	13	$v : b$	9	24	$v : v_{pt}$	3	35	$p : p$	1
3	$v : v$	141	14	$n : n_s$	8	25	$v_g : v_s$	3	36	$p : c$	1
4	$n : a$	97	15	$v : i$	7	26	$a : b$	3	37	$p : o$	1
5	$v : a$	34	16	$n : v_s$	6	27	$n : c$	2	38	$c : t$	1
6	$n_s : n_s$	30	17	$n : v_{pt}$	6	28	$n : o$	2	39	$n : v_{pt}$	1
7	$n : b$	24	18	$v_d : a$	6	29	$b : p$	2	40	$n : v_{pt}$	1
8	$n : i$	20	19	$b : b$	4	30	$p : t$	2	41	$n : a_r$	1
9	$n_s : v_s$	19	20	$n : v_{pt}$	4	31	$v_{ng} : p$	2	42	$v : v_{pt}$	1
10	$a : a$	17	21	$v : v_{pt}$	4	32	$v_{pt} : v_{pt}$	2	43	$v : a_r$	1
11	$n : v_{ng}$	11	22	$v_{ng} : v_{ng}$	4	33	$n : t$	1	44	$v_{ng} : a$	1
										Всего	1627

на этом уровне меньше: из 24 типов уникальными являются только 5, из 44 редуцированных — 12.

В таблицах 20 и 21 представлено распределение гиперлексем по частям речи (для простых ГЛ) и сложным лексико-грамматическим разрядам (для сложных ГЛ). Под сложным лексико-грамматическим разрядом гиперлексемы понимается совокупность лексико-грамматических классов (частей речи), к которым принадлежат входящие в состав гиперлексем слова (чисто грамматические, или моделированные, омонимы). Выявление принадлежности гиперлексем к различным частям речи и сложным лексико-грамматическим разрядам представляет большой интерес не только для изучения омонимии, но и для лексикологических исследований вообще, поскольку оно дает подробную структурно-квантитативную характеристику конверсионных отношений между словами на определенном участке словарного состава английского языка.

Таблица 20

Распределение гиперлексем по простым и сложным лексико-грамматическим разрядам

№ разряда	Разряд	Количество ГЛ	№ разряда	Разряд	Количество ГЛ	№ разряда	Разряд	Количество ГЛ	№ разряда	Разряд	Количество ГЛ
1	*n*	624	8	*b*	6	15	*vab*	2	22	*nvbpco*	1
2	*nv*	251	9	*i*	6	16	*bp*	2	23	*vb*	1
3	*v*	123	10	*nvi*	4	17	*t*	2	24	*vp*	1
4	*a*	57	11	*p*	4	18	*nvbt*	1	25	*ai*	1
5	*na*	22	12	*nvb*	3	19	*ni*	1	26	*bco*	1
6	*nva*	13	13	*nvab*	3	20	*nvabp*	1	27	*c*	1
7	*va*	8	14	*ab*	3	21	*nvp*	1			
										Всего	1143

Таблица 21

Распределение гиперлексем по лексико-грамматическим разрядам с учетом внутренних грамматических значений

№ разряда	Разряд	Количество ГЛ	№ разряда	Разряд	Количество ГЛ	№ разряда	Разряд	Количество ГЛ	№ разряда	Разряд	Количество ГЛ
1	*n*	600	11	*b*	6	21	*bp*	2	31	*vb*	1
2	*nv*	237	12	*i*	6	22	*t*	2	32	*ai*	1
3	*v*	100	13	*nvi*	4	23	nv_{ng}	2	33	*bco*	1
4	*a*	56	14	*va*	4	24	$v_{ng}a$	2	34	*c*	1
5	n_g	24	15	*p*	4	25	$v_{pt}a$	2	35	nv_{pt}	1
6	*na*	22	16	v_g	4	26	*nvbi*	1	36	$nv_{pt}a$	1
7	*nva*	12	17	*nvb*	3	27	*ni*	1	37	nva_rb_r	1
8	n_sv_s	11	18	*ab*	3	28	*nvabp*	1	38	$v_{ng}p$	1
9	v_{pt}	10	19	*nvab*	2	29	*nvp*	1	39	v_{pt}	1
10	v_{ng}	8	20	*vab*	2	30	*nvbpco*	1	40	a_r	1
										Всего	1143

Из таблиц 14—17 и 20—21 видно, что разряды ГЛ отличаются меньшим разнообразием, чем типы ОГ: здесь всего 27 разрядов, выделенных с учетом «внешних», категориальных, грамматических значений, и 40 разрядов, выделенных с учетом внутренних значений. Доля уникальных разрядов также не столь велика — около 37% общего числа разрядов; однако наиболее частотные разряды чрезвычайно емки: уже первые три, образованные субстантивными и глагольными ГЛ (n, nv и v), охватывают почти 90% всех ГЛ. Таким образом, и в сфере чисто грамматической омонимии субстантивные и глагольные единицы преобладают.

Наблюдается некоторое сходство между распределением разрядов ГЛ (табл. 20) и редуцированных структурных типов ОГ (табл. 16). Однако сходство это чисто внешнее, вызванное сходством формул рассматриваемых разрядов и типов. Формула редуцированного типа дает весьма упрощенное отражение структуры ОГ, так как опускает обозначения повторяющихся частей речи; формула же ГЛ отражает действительную структуру ГЛ, где слова одной и той же части речи не повторяются (по определению). Таким образом, и в том, и в другом случае символ каждой части речи может присутствовать в формуле лишь однажды, что и обусловливает сходство формул. Различие же между этими двумя распределениями состоит прежде всего в том, что если из 27 типов ОГ только 4 не содержат в своем составе существительных или глаголов, то из 27 разрядов ГЛ безглагольных и безсубстантивных насчитывается 10. Другое различие заключается в том, что среди типов ОГ имеется только 4 простых (n, v, a и b), т. е. простые ОГ могут быть только субстантивными, глагольными, адъективными и адвербиальными, тогда как простые ГЛ могут быть также и междометными, предложными и т. д. (типы i, p, c, t). Аналогичным образом различаются также разряды ГЛ и редуцированные типы ОГ, выделенные с учетом внутренних грамматических значений (табл. 17 и 21).

Таблица 22

Соотношение количества простых и сложных омогрупп, гиперлексем и межлексемных омопар

Структурные единицы	Количество ОГ		Количество ГЛ		Количество межлексемных омопар	
	в абсолютных цифрах	в %%	в абсолютных цифрах	в %%	в абсолютных цифрах	в %%
Простые	167	34.8	823	72.0	846	52.0
Сложные	313	65.2	320	28.0	781	48.0
Всего	480	100.0	1143	100.0	1627	100.0

Соотношение числа простых и сложных ОГ, ГЛ и межлексемных омопар представлено в таблице 22. Как видно из таблицы, среди омогрупп простых единиц почти вдвое меньше, чем сложных, среди гиперлексем, наоборот, простых в 2,5 раза больше, чем сложных, а среди межлексемных пар омонимов простых пар встречается примерно столько же, сколько и сложных. Эти цифры характеризуют соотношение чисто лексической («простой») и лексико-грамматической («сложной») омонимии на разных структурных уровнях.

Таблица 23

Распределение омонимов по частям речи

№ п/п	Части речи	Количество омонимов	
		в абсолютных цифрах	в %%
1	n	925	61.6
2	v	413	27.5
3	a	110	7.3
4	b	24	1.6
5	i	13	0.9
6	p	10	0.7
7	c	3	0.2
8	o	2	0.1
9	t	2	0.1
Всего		1502	100.0

Таблица 24

Распределение омонимов по лексико-грамматическим классам и подклассам

№ п/п	Классы и подклассы	Количество омонимов	
		в абсолютных цифрах	в %%
1	n	890	59.3
2	v	370	24.7
3	a	108	7.2
4	n_g	35	2.3
5	b	23	1.5
6	v_g	15	1.0
7	v_{pt}	14	0.9
8	i	13	0.9
9	v_{ng}	13	0.9
10	p	10	0.7
11	c	3	0.2
12	o	2	0.1
13	t	2	0.1
14	a_r	2	0.1
15	v_{pt}	1	0.05
16	b_r	1	0.05
Всего		1502	100.0

Анализ соотношения структурных типов омонимических единиц разных уровней завершают таблицы 23 и 24, где представлено распределение омонимов по лексико-грамматическим классам (частям речи) и подклассам. Омонимы-существительные составляют в изученном материале около 60% общего числа омонимов, глаголы — более одной четверти, а на долю знаменательных частей речи вместе приходится 99% всех омонимов. Эти цифры близки к данным, полученным на материале фонетических омонимов (см.: Тышлер, 1975, с. 368), отличаясь от них лишь несколько более низким процентом глаголов.

Выявление структурного состава исследованных омонимических единиц позволило определить распространенность грамматической дифференциации на разных уровнях. Выяснилось, что грамматически дифференцированные омонимы составляют 26%

Таблица 25

Роль грамматической дифференциации омонимов на уровне слов и на уровне словоформ

Лексический уровень	Уровень омонимических единиц	Количество единиц			Дифференцированные единицы в %% к общему числу единиц данного вида
		всего	в том числе:		
			недифференцированных	дифференцированных	
Словоформы	омоблоки (РСФ)	3226	2588	638	19.8
	формемы	2527	2280	247	9.8
	омокомплекты	1096	992	104	9.5
Слова	омонимы	1502	1105	397	26.4
	гиперлексемы	1143	945	198	17.3
	омогруппы	480	405	75	15.6

общего числа омонимов (см. табл. 25). Это означает, что на уровне слов лексическая неоднозначность омонимов в 26% случаев может быть снята средствами грамматического анализа. На уровне словоформ (точнее, блоков словоформ) эта цифра составляет около 20%. Важно отметить, что более 15% общего числа омогрупп состоят целиком из грамматически дифференцированных омонимов.

6. Распределение омонимов по длине слова

В ряде работ по омонимии затрагивается вопрос о длине омонимов (см.: Jespersen, 1928, с. 356; Ullmann, 1964; с. 78; и др.). Высказывается предположение о том, что среди омонимов процент односложных слов выше, чем в языке в целом: 80% по сравнению с 58% среди всех слов языка (Тышлер, 1975, с. 73). По нашим данным, процент односложных слов среди омонимов значительно ниже. Из 476 словарных ОК односложными являются всего 252, т. е. 53%. По-видимому, среди графических омонимов односложных слов значительно меньше, чем среди омонимов фонетических, которые подсчитывались в работе И. С. Тышлера.

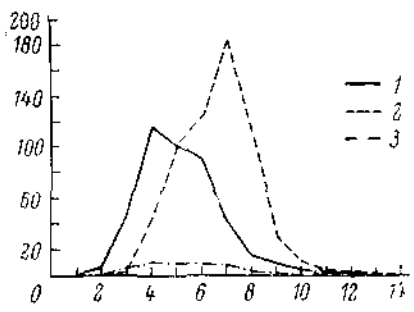


Рис. 10. Распределение словарных (1), несловарных (2) и смешанных (3) омокомплектов по длине словоформ.

По оси ординат — количество ОК, по оси абсцисс — длина омоформ (в буквах).

Представляет интерес распределение графических омонимов (омоформ) по длине слова (словоформы) в буквах (см. табл. 26). Бросается в глаза различие в длине словарных и несловарных омоформ: словарные омоформы чаще всего состоят из четырех букв, смешанные — из пяти, а несловарные — из семи (см. рис. 10)

Таблица 26

Распределение словарных, смешанных и несловарных омокомплектов по длине словоформы (в буквах)

Длина словоформы	Количество ОК с данной длиной словоформы			
	словарных	смешанных	несловарных	всего
1	2	—	—	2
2	6	—	—	6
3	46	5	1	52
4	116	8	45	169
5	101	9	104	214
6	90	7	126	223
7	43	6	183	232
8	15	3	110	128
9	9	2	31	42
10	4	—	10	14
11	2	—	4	6
12	2	—	3	5
13	—	—	2	2
14	—	—	1	1
Всего	436	40	614	1096
В том числе:				
а) 1 — 5 букв	271 (62.2%)	22 (55.0%)	150 (24.4%)	443 (40.4%)
б) 6 и более	165 (37.8%)	18 (45.0%)	464 (75.6%)	653 (59.6%)
Средняя длина словоформы	5.16	5.45	6.20	6.01

Коротких словоформ (до 5 букв включительно, т. е. преимущественно односложных) среди словарных ОК почти вдвое больше, чем длинных, тогда как среди несловарных ОК положение обратное: длинных омоформ втрое больше, чем коротких. В среднем несловарная омоформа длиннее словарной на одну букву, что объясняется наличием у нее словоизменительного аффикса, самый распространенный из которых является однобуквенным (-s). Для выяснения вопроса о сравнительной длине омонимичных и неомонимичных словоформ необходимо дополнительное исследование.

7. Омонимическая насыщенность словарного состава

Большой теоретический и практический интерес представляет вопрос об омонимической насыщенности словаря, т. е. о доле омонимов среди всех слов языка. Попытки определить количество омонимов в английском языке предпринимались неоднократно (см.: Bridges, 1919, с. 6; Костюченко, 1954; Тышлер, 1966; 1975), но, вследствие нечеткости методологических установок авторов, полученные результаты оказались противоречивыми. При этом никто из исследователей, за исключением А. Я. Шайкевича (Шайкевич, 1962), не пытался сопоставить число выявленных омонимов с общим числом слов в словаре. Однако наши результаты несопоставимы с данными А. Я. Шайкевича, потому что в его работе подсчитывались не графические, а фонетические омонимы и не учитывались омонимы лексико-грамматические. Наиболее корректно определяется «коэффициент омонимичности» в работе П. М. Алексеева (Алексеев, 1965, с. 298), который, как и мы, рассматривал омонимы графические, как чисто лексические, так и лексико-грамматические. Однако и в этом случае полученные результаты несопоставимы с нашими, поскольку в указанной работе единицей счета служили не слова, а словоформы, и подсчет производился на основе словаря узкоспециального и меньшего объема.

Для того чтобы определить долю омонимов во всем словаре (или в определенном отрезке словаря), нужно знать общее количество слов в этом словаре или отрезке. В АРСл каждая словарная статья соответствует гиперлексеме (сложной или простой), а каждый раздел статьи, обозначенный жирной арабской цифрой — отдельному слову. Это дает возможность определить омонимическую насыщенность словаря как на уровне слов, так и на уровне гиперлексем. В COD соответствия между количеством статей и количеством гиперлексем, так же как и между количеством разделов и количеством слов, нет; поэтому в этом словаре нам пришлось подсчитывать все слова подряд — как заголовочные, так и приводимые внутри статьи. Оценка общего числа слов и гиперлексем в отрезке A—E обоих словарей производилась методом случайной выборки с доверительным уровнем 95%. На этой основе была определена омонимическая насыщенность словаря на уровне гиперлексем для АРСл и на уровне слов — для АРСл и COD (см. табл. 27). Исходя из данных таблицы 27, можно ожидать, что на материале всего словаря (A—Z) общее количество гиперлексем составит около 4000, а омонимов — более 5000.

Полученные результаты (7—9% омонимичных слов и 8% омонимичных гиперлексем) не могут быть пока расценены как показатель сильной или, наоборот, слабой насыщенности английского языка омонимами, поскольку сопоставимых цифр по другим языкам не имеется. По мнению многих авторов (Костюченко, 1954, с. 40; Воронцова, 1960, с. 20; Эман, 1960, с. 118; Ullmann, 1964,

Таблица 27

Омонимическая насыщенность словарного состава (ОНСС)

Словарь	Количество гиперлексем в отрезке A—E		ОНСС на уровне гиперлексем	Количество слов в отрезке A—E		ОНСС на уровне слов
	всего	из них омонимичных		всего	из них омонимичных	
АРСл	13700±1100	1143	8.1%±0.7%	16550±1250	1502	9.1%±0.7%
COD	—	—	—	19250±1050	1502	7.8%±0.4%

Примечание. Количество гиперлексем в АРСл принимается равным количеству словарных статей (за вычетом отсылочных статей и статей на аффиксы и словосочетания). Количество гиперлексем в COD, где слова не сгруппированы в гиперлексемы, не определялось.

с. 78; и др.), не подкрепляемому, правда, никакими расчетами, английский язык более богат омонимами, чем многие другие языки, в том числе немецкий, французский и русский. Проверить это мнение до настоящего времени не представлялось возможным из-за неразработанности методики сравнительно-типологического изучения омонимии (см. об этом также: Малаховский, 1972, с. 272; 1977, с. 94). Изложенная в настоящей статье методика выделения и количественного определения омонимических единиц разных уровней может послужить базой для подобного рода исследований и, в частности, осуществить сопоставление величины омонимической насыщенности словаря в разных языках.

8. Распределение омонимических рядов по частоте

В ряде работ по омонимии делается попытка выяснить, относятся омонимы к высокочастотным или к низкочастотным слоям лексики. Мнения исследователей расходятся: одни считают, что среди наиболее употребительных слов языка доля лексических омонимов мала и что она возрастает с уменьшением частоты слова (Шайкевич, 1962, с. 291), другие — что она, наоборот, возрастает по мере приближения к ядру словарного состава (Тышлер, 1975, с. 73). По данным П. М. Алексеева, большинство омонимов являются редко употребляемыми словами (Алексеев, 1965, с. 298).

Для решения этого вопроса необходимо прежде всего установить, какие слова мы считаем частыми, а какие — редкими. В настоящем исследовании была использована методика разбиения словарного состава языка на частотные зоны, изложенная на с. 99—105 данного сборника.

Распределение омогрупп по частотным зонам, полученное на основе данных словаря Торндайка-Лорджа (Thorndike, Lorge,

Таблица 28

Распределение омогрупп по частотным зонам (на основе данных словаря Торндайка-Лорджа)

Частотная зона	Количество ОГ в зоне	
	в абсолютных цифрах	в %%
A	1	0.2
B	11	2.3
C	36	7.5
D	156	32.5
E	152	31.7
F	47	9.8
Всего в этих зонах	403	84.0
Нет в словаре	77	16.0
Всего ОГ	480	100.0

1944), представлено в таблице 28. Из таблицы видно, что бо́льшая часть омогрупп относится к зонам D и E, т. е. к зонам средних по частоте и редких слов; в зонах частых слов (A, B, C) насчитывается всего 10% омогрупп, а в зоне очень редких (F) также около 10%. Более наглядно распределение омогрупп по частоте показано на рисунке 11, где каждая зона разбита на два участка (числовые характеристики зон даны в таблице 2 на с. 103).

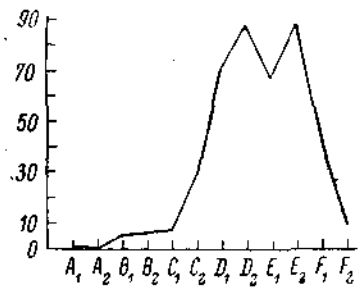


Рис. 11. Распределение омогрупп по частотным зонам (по данным словаря Торндайка-Лорджа).

По оси ординат — количество ОГ, по оси абсцисс — частотные зоны.

Следует учесть, что распределение омогрупп по частотным зонам словаря Торндайка-Лорджа является не вполне корректной процедурой, так как каждая омогруппа состоит из двух или нескольких гиперлексем (т. е. из нескольких слов), тогда как единицей счета в данном словаре являются «слова», в качестве которых выступают либо отдельные лексемы, либо гиперлексемы, а иногда и омогруппы. Мы вынуждены прибегнуть к этой процедуре за неимением других возможностей выявить картину частотного распределения омонимов на уровне слов. На уровне словоформ такая возможность имеется, поскольку современные частотные словари используют в качестве единицы счета «словоформы»,

Таблица 29

Распределение омокомплектов по частотным зонам (на основе данных словаря Кэрролла)

Частотная зона	Количество ОК	
	в абсолютных цифрах	в %%
A	1	0.1
B	12	1.1
C	17	1.5
D	161	14.7
E	265	24.2
F	186	17.0
G	126	11.5
Всего в частотных зонах	768	70.1
Вне этих зон	328	29.9
Всего ОК	1096	100.0

к числу которых они относят и группы омонимичных словоформ, т. е. по нашей терминологии омокомплекты.

Распределение омокомплектов по частоте (см. табл. 29) произведено на основе частотного словаря Кэрролла. При этом использовалась не фактически наблюдавшаяся частота (F), а величина U, характеризующая вероятность появления словоформы в словаре неограниченно большого объема и получаемая путем преобразования частоты словоформы с учетом ее дисперсии. Как видно из таблицы, максимальное число омокомплектов относится к зоне E, т. е. к зоне редких словоформ, а довольно значительное их число находится также в зонах D, F и G.

Таблицы 28 и 29 дают ответ на вопрос о том, к каким словам относятся омонимы — к частым или редким. На уровне слов большинство омонимов — это слова средние по частоте ($F=10 \div 99$ миллионных) или редкие ($F=1 \div 9$ миллионных); на уровне словоформ — это формы, главным образом, редкие и очень редкие ($U=0.1 \div 0.9$ миллионных). Однако из этих таблиц нельзя узнать, какова доля омонимов и омоформ среди частых и редких слов языка и отличается ли частотное распределение омонимических единиц от распределения всех слов и словоформ языка. Дать точный ответ на этот вопрос в настоящее время невозможно; однако есть возможность получить приблизительную оценку интересующих нас величин.

В таблице 30 дается оценка частотного распределения омокомплектов во всем словаре (A—Z), полученная из их фактического

Таблица 30

Оценка частотного распределения омокомплектов во всем словаре и их доли в общем количестве словоформ языка

Частотная зона	Количество ОК в первой трети словаря (отрезок от A до E)	Оценка количества ОК во всем словаре (A—Z)	Общее количество словоформ в словаре	Оценка доли ОК в общем количестве словоформ (в %%)
A_1	0	0	2	0.0
A_2	1	3	6	50.0
B_1	6	20	26	76.7
B_2	4	13	58	22.4
C_1	9	30	199	15.1
C_2	15	50	779	6.4
D_1	58	193	2345	8.2
D_2	121	403	5069	8.0
E_1	154	514	8242	6.2
E_2	245	817	34 607	2.4
Всего в зонах A_1—E_2	613	2043	51 323	4.0
Вне зон	483	1610	—	—
Всего ОК	1096	3653	—	—

распределения в отрезке словаря от A до E (составляющем около 30% общего объема словаря) на основе допущения о равномерном распределении омокомплектов по всему словарю. Судя по оценке, доля омокомплектов в общем числе словоформ особенно велика в высокочастотных зонах (A и B) и постепенно снижается при переходе к низкочастотным зонам. В связи с тем, что большинство омонимов — низкочастотные слова, суммарная оценка доли ОК в общем числе словоформ оказалась также низкой.

Таким образом, частотное распределение омокомплектов характеризуется двумя разнонаправленными тенденциями: с одной стороны, общее число ОК при переходе от частых словоформ к редким все более увеличивается, с другой стороны, доля омонимичных ОК в общем числе словоформ при этом неуклонно снижается.

9. Зависимость изученных количественных характеристик от частоты омонимических единиц

Использование описанной выше (с. 99—105) методики частотной стратификации лексики позволяет уточнить и углубить решение ряда задач, уже рассматривавшихся нами без обращения к частотной характеристике исследуемых омонимических единиц.

Таблица 31

Зависимость объема ОК и формемы от частоты ОК (по данным словаря Кэрролла)

Частотная зона	Количество ОК в зоне	Количество формем		Количество омоблоков		
		всего	в среднем в одном ОК	всего	в среднем в одной формеме	в среднем в одном ОК
A	1	4	4.00	4	1.00	4.00
B	12	45	3.75	50	1.11	4.17
C	17	43	2.53	69	1.60	4.06
D	161	401	2.49	592	1.48	3.68
E	265	650	2.45	886	1.36	3.34
F	186	423	2.27	521	1.23	2.80
G	126	276	2.19	319	1.16	2.53
Всего в зонах A—G	768	1842	2.40	2441	1.33	3.18
Вне этих зон	328	705	2.15	787	1.12	2.40
Всего	1096	2547	2.32	3228	1.27	2.95

Изучение зависимости объема омокомплекта и формемы от частоты ОК показало (см. табл. 31), что среди высокочастотных омокомплектов больше многочленных и имеющих развитую, т. е. глагольную, парадигму. Чем выше частота ОК, тем больше в нем формем, т. е. тем сильнее развита лексическая омонимия. С другой стороны, чем выше частота ОК, тем больше он содержит омоблоков, т. е. тем сильнее развита внутренняя грамматическая омонимия, характерная для глагола. По мере продвижения к низкочастотным зонам количество формем и количество омоблоков все больше приближается к двум — т. е. к простой паре субстантивных словоформ ($n : n$ или $n_s : n_s$).

Изучение зависимости грамматической дифференциации омонимических единиц от частоты показало (табл. 32), что в высокочастотных зонах грамматическая дифференциация омонимов развита сильнее, чем в низкочастотных, что особенно хорошо заметно в частотном ряду омоблоков. Это говорит о том, что роль грамматической дифференциации в разграничении омонимов фактически выше, чем можно было предположить по данным таблицы 25.

Наконец, была выявлена зависимость между длиной омонимов и частотой омонимического ряда — как на уровне слов (табл. 33), так и на уровне словоформ (табл. 34). Выяснилось, что длина омонимичного слова (представленного его словарной формой), так же как и длина любой омоформы, находится в обратной зависимости от частоты омонимического ряда.

Таблица 32

Распространенность грамматической дифференциации омокомплектов, формем и омоблоков в разных частотных зонах

Частотная зона	Количество ОК			Количество формем			Количество омоблоков		
	всего	в том числе грамматически дифференцированных		всего	в том числе грамматически дифференцированных		всего	в том числе грамматически дифференцированных	
		в абсолютных цифрах	в ‰		в абсолютных цифрах	в ‰		в абсолютных цифрах	в ‰
A	1	1	100.0	4	4	100.0	4	4	100.0
B	12	5	41.7	45	11	24.4	50	30	60.0
C	17	3	17.6	43	8	18.6	69	23	33.3
D	161	16	9.9	401	42	10.4	592	133	22.5
E	265	28	10.6	650	60	9.2	886	174	19.6
F	186	8	4.3	423	32	7.6	521	105	20.2
G	126	16	12.7	276	33	12.0	319	61	19.1
Всего в зонах A—G	768	77	10.0	1842	190	10.3	2441	530	21.7
Вне этих зон	328	27	8.2	685	57	8.3	785	108	13.8
Всего	1096	104	9.5	2527	247	9.7	3226	638	19.8

Заключение: Омонимика и системность

Структурные и квантитативные характеристики омонимических рядов и омонимических единиц английского языка, полученные в настоящем исследовании, могут оказаться полезными для решения различных задач ИЛ, в частности — для определения объема АС. Эти данные могут иметь и определенное теоретическое значение, поскольку они убедительно показывают, что представление об омонимике языка как о бессистемной, неорганизованной совокупности никак не связанных между собой языковых единиц, случайно совпавших по звучанию или написанию, является неверным. В английской омонимике прослеживаются определенные системные связи, которые на разных уровнях носят разный характер.

На уровне омогрупп это связи деривационного характера, соответствующие связям между кодериватами при словообразовании. Некоторые омогруппы объединяются в своего рода гроздья (по две — три омогруппы в каждой), и внутри каждого такого

Таблица 33

Зависимость между длиной омонимов и частотой омогруппы (по словарю Торндайка—Лорджа)

Длина слова в буквах	Количество ОГ в частотных зонах													Нерегистрируемых форм	Итого
	A_1	A_2	B_1	B_2	C_1	C_2	D_1	D_2	E_1	E_2	F_1	F_2	H	Всего	
1	1	—	—	—	—	—	—	—	—	—	—	—	1	2	—	2
2	—	—	3	2	—	—	—	—	—	—	—	—	1	6	—	6
3	—	—	2	1	2	9	8	9	6	11	1	—	2	51	—	51
4	—	—	—	2	3	12	25	35	17	16	5	—	8	123	3	126
5	—	—	—	1	1	4	19	15	24	17	10	2	10	103	7	110
6	—	—	—	—	1	3	8	16	7	26	16	2	14	93	5	98
7	—	—	—	—	—	1	6	9	7	9	6	1	8	47	3	50
8	—	—	—	—	—	—	1	2	3	4	—	1	6	17	1	18
9	—	—	—	—	—	—	3	—	1	3	—	1	3	11	—	11
10	—	—	—	—	—	—	—	—	—	1	1	—	2	4	—	4
11	—	—	—	—	—	—	—	—	—	—	—	—	2	2	—	2
12	—	—	—	—	—	—	—	—	—	—	1	—	1	3	—	2
Всего	1	—	5	6	7	29	70	86	65	87	39	8	58	461	19	480
Средняя длина слова	1	—	2.4	3.3	4.1	4.1	4.9	4.8	5.1	5.4	5.9	6.1	6.2	5.17	5.6	5.18

Примечания. 1. Учитывалась, как правило, длина словарной формы слова; в смешанных омогруппах, основанных на омонимии несловарных форм, учитывалась длина несловарной формы. 2. Зона F_2 охватывается словарем Торндайка-Лорджа не полностью (она представлена в словаре только частотами 0.22 и 0.28; см. примечание 2 к табл. 2 на с. 103).

Таблица 34

Зависимость между длиной омоформы и частотой омокомплекта (по словарю Кэрролла)

Длина омоформы в буквах	Количество ОК в частотной зоне														Всего
	A_1	A_2	B_1	B_2	C_1	C_2	D_1	D_2	E_1	E_2	F_1	F_2	G_1	G_2	H	
1	—	1	—	—	—	—	1	—	—	—	—	—	—	—	—	2
2	—	—	3	2	—	—	—	—	—	—	1	—	—	—	—	6
3	—	—	2	3	2	3	11	4	5	6	8	5	1	1	1	52
4	—	—	1	2	1	7	26	25	24	21	12	14	2	7	28	169
5	—	—	—	1	1	2	18	22	36	38	12	18	5	16	46	214
6	—	—	—	—	—	1	10	13	26	26	28	28	8	22	61	223
7	—	—	—	—	—	—	5	16	29	23	24	13	12	26	84	232
8	—	—	—	—	—	—	2	4	9	12	10	4	3	18	66	128
9	—	—	—	—	—	—	1	3	3	4	2	4	—	2	23	42
10	—	—	—	—	—	—	—	—	1	1	1	—	—	1	10	14
11	—	—	—	—	—	—	—	—	1	—	1	—	—	1	3	6
12	—	—	—	—	—	—	—	—	—	—	—	1	—	1	3	5
13	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2	2
14	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1
Всего омокомплектов	—	1	5	7	4	13	74	87	134	131	99	87	31	95	328	1096
Средняя длина омоформы	—	1	2.4	3.0	3.8	4.1	4.7	5.4	5.7	5.7	5.9	5.8	6.2	6.5	6.8	6.0

объединения словообразовательные отношения, а следовательно, и отношения сходства лексических значений, между членами различных омогрупп являются регулярными, повторяющимися. *Системные связи такого рода охватывают более 15% омонимики на* этом уровне.

На уровне омокомплектов это деривационные связи, соответствующие связям между членами одной парадигмы при формообразовании. Многие омокомплекты объединяются в группы (по два — *четыре омокомплекта в омогруппе), и внутри каждой такой группы* отношения формообразования между членами разных омокомплектов являются регулярными и в большой мере предсказуемыми. Системные связи этого рода охватывают свыше 90% омонимики на данном уровне.

На уровне слов (омонимов) системные связи носят двоякий характер. С одной стороны, омонимы, различающиеся по лексическому значению и, следовательно, относящиеся к разным гиперлексемам, могут принадлежать одной и той же части речи, т. е. образовывать внутри одной омогруппы грамматически однородные *пары. Таким образом, возникают регулярные* грамматические связи между членами разных гиперлексем в пределах одной омогруппы. Подобные грамматические связи наблюдаются в большинстве случаев лексической омонимии: грамматически однородными являются около 55% общего числа пар лексических омонимов. *С другой стороны, омонимы, входящие в одну и ту же гиперлексему*, связаны между собой не просто общностью лексического значения и тождеством формы, но и общностью корня. Таким образом, гиперлексема образует внутри омогруппы четкую подсистему слов, находящихся в отношениях словообразования. *Отношения эти регулярно воспроизводятся во многих других* гиперлексемах языка, благодаря чему за омонимией этого типа закрепилось название «моделированной». Системные отношения этого рода охватывают около 30% слов в нашем материале, т. е. в составе изученных нами лексических омогрупп; приблизительно *такова же, вероятно, доля моделированных омонимов и среди* всех слов языка вообще (применительно к словарю объемом в 60—80 тысяч слов).

Наконец, на уровне омоформ проявления системности носят еще более многообразный характер. Грамматические связи внутри *омокомплекта* устанавливаются как между *омоформами*, грамматически однородными (с тождеством категориальных грамматических значений), так и между омоформами, грамматически *тождественными (т. е. с тождеством и внутренних грамматических* значений). Лексико-семантической связью охвачены все *омоформы одной формемы*. Омоформы связаны между собой и деривационными отношениями словообразовательного характера, но эти отношения вытекают из отношений языковых единиц более высоких уровней и на данном уровне в специальном рассмотрении не нуждаются.

Важно отметить, что многие омоформы в составе омокомплекта связаны между собой как отношениями грамматической однородности, так и отношениями сходства лексических значений; это формы, принадлежащие парадигме одного и того же слова. Связь между ними так велика, что в обычной лингвистической практике они рассматриваются как единая словоформа; в настоящей же работе такие группы тесно спаянных омоформ квалифицируются как омоблоки. Количество словоформ, организующихся в блоки по 2—3 и более форм в каждом, составляет в изученном материале не менее 60% общего числа омоформ.

Таким образом, омонимика языка представляет собой некоторую слабо организованную систему,[16] отдельные участки и уровни которой в большей или меньшей степени пронизаны отношениями системности. Сильнее всего эти отношения проявляются на нижних уровнях (на уровне омоформ и уровне слов) и на таком обширном участке, как чисто грамматическая (моделированная) омонимия, слабее — на уровнях омокомплектов и омогрупп и на участке омонимии лексической. Специального исследования заслуживает вопрос о тождестве формы как о факторе системности в омонимике и о противоречии между тождеством формы и различием значения как о движущей силе развития омонимики в ходе эволюции языка.

[16] О различии между слабой (неорганичной) системой и системой сильной (органичной) см.: Блауберг, Юдин, 1973, с. 176; Малаховский, 1975, с. 188.

ЛИТЕРАТУРА

Абаев В. И. О подаче омонимов в словаре. — ВЯ, 1957, № 3.

Автоматическая переработка текста. (Системный подход в решению проблемы машинного перевода). Кишинев, 1978.

Адмони В. Г. Основы теории грамматики. М.—Л., 1964.

Адмони В. Г. Еще раз об изучении количественной стороны грамматических явлений. — ВЯ, 1970, № 1.

Александрова З. Е. Словарь синонимов русского языка. Около 9000 синонимических рядов. Изд. 4-е. М., 1975.

Алексеев П. М. Частотный словарь английского подъязыка электроники. Автореф. канд. дис. Л., 1965.

Алексеев П. М. Частотный англо-русский словарь-минимум по электронике. Л., 1971.

Алексеев П. М., Турыгина Л. А. Частотный англо-русский словарь-минимум газетной лексики. М., 1974.

Амосова Н. Н. Этимологические основы словарного состава английского языка. М., 1956.

Апресян Ю. Д. Лингвистическая семантика. Синонимические средства языка. М., 1974.

Арзикулов Х. А., Пиотровский Р. Г., Попеску А. Н., Хажинская М. С. Автоматизированная система тезаурусного аннотирования научно-технического документа. — НТИ, серия 2, 1978, № 12.

Арнольд И. В. Семантическая структура слова в современном английском языке и методика ее исследования. (На материале имени существительного). Л., 1966.

Арнольд И. В. Лексикология современного английского языка. Изд. 2-е. М., 1973.

Ахманова О. С. Очерки по общей и русской лексикологии. М., 1957.

Ахманова О. С. Словарь лингвистических терминов. М., 1966.

Ахманова О. С. Вопрос о языковой картине мира и проблема омонимии. — L'Annuaire de l'Institut de Philologie et d'Histoire Orientales et Slaves, t. XVIII (1966—1967). Bruxelles, 1968.

Бектаев К. Б. Статистико-информационная типология тюркского текста. Алма-Ата, 1978.

Билан В. Н. Об одном подходе к формализации семантики. — В кн.: Вопросы общей и прикладной лингвистики (сборник теоретических статей). Минск, 1975.

Беляева Л. Н., Лукьянова Е. М., Пиотровский Р. Г. Машинная лексикография в многоцелевом автоматическом русском словаре. — В кн.: Лингвистические исследования, 1976. Вопросы лексикологии, лексикографии и прикладной лингвистики. М., 1976.

Блауберг И. В., Юдин Э. Г. Становление и сущность системного подхода. М., 1973.

Бондарко А. В. Вид и время русского глагола. М., 1971.

Борисевич А. Д. и др. Кодирование грамматической информации в машинном словаре. — В кн.: Статистика текста. Минск, 1970.

Булаховский Л. А. Из жизни омонимов. — Русская речь, новая серия, вып. 3, 1928.

Васильева М. И. Семантико-синтаксический анализ адъективно-именных словосочетаний (на материале общественно-политических текстов). Автореф. канд. дис. Л., 1978.

Венецкий И. Г., Кильдишев Г. С. Теория вероятностей и математическая статистика. Изд. 3-е. М., 1975.

Вертель В. А., Вертель Е. В., Крисевич В. С., Пиотровский Р. Г., Трибис Л. Н. Автоматические словари в системе бинарного вероятностного МП. — Учен. зап. Ленинградского пед. ин-та им. А. И. Герцена, т. 458, ч. 1. Л., 1971.

Виноградов В. В. О грамматической омонимии в современном русском языке. — РЯШ, 1940, № 1.

Войнов В. К. Количественный анализ текста для описания индивидуального стиля. Автореф. канд. дис. Киев, 1972.

Вольская И. С. Дифференциальные признаки официально-делового стиля речи на синтаксическом уровне. — Учен. зап. 1-го Московского пед. ин-та иностр. языков им. М. Тореза, т. 33. М., 1965.

Воронцова Г. Н. Очерки по грамматике английского языка. М., 1960.

Головин Б. Н. Язык и статистика. М., 1971.

Гончаренко В. В., Котельникова Н. М. Автоматический синтаксический анализ двусоставных английских предложений. — В кн.: Романское языкознание. Вып. 1. Вопросы экспериментальной фонетики и прикладной лингвистики. Минск, 1978.

Горбунов Ю. И. Формальное распознавание значений предлогов французского языка. Автореф. канд. дис. Л., 1978.

Горевая В. С. Абсолютные и дифференциальные признаки некоторых стилевых рубрик английского языка. — Учен. зап. Калининского пед. ин-та, т. 64, вып. 1, ч. 2. Калинин, 1969.

Грехнева Г. М. Стилевая дифференциация придаточных предложений в современном русском языке. — В кн.: ВСС, 1974.

Единая система ЭВМ. М., 1974.

Жигадло В. Н., Иванова И. П., Иофик Л. Л. Современный английский язык. М., 1956.

Журавлев А. П. Опыт вероятностно-статистического изучения стилевых различий. — В кн.: Язык и общество. Саратов, 1967.

Журавлев А. П. К статистическому изучению функциональных стилей. — Учен. зап. Калининградского ун-та. Филол. науки, вып. 4. Калининград, 1969.

Задорожный М. И. О границах полисемии и омонимии. М., 1971.

Закс Л. Статистическое оценивание. М., 1976.

Зиндер Л. Р., Строева Т. В. К вопросу о применении статистики в языковедении. — ВЯ, 1968, № 6.

Иванова И. П. О морфологической характеристике слова в современном английском языке. — В кн.: Проблемы морфологического строя германских языков. М., 1963.

Информационные системы общего назначения. М., 1975.

История римской литературы. Т. 1. М., 1959.

Кауфман А. А. Теория и методы статистики. Изд. 5-е. М.—Л., 1928.

Кауфман С. И. Из курса лекций по статистической стилистике. М., 1970.

Кацнельсон С. Д. Типология языка и речевое мышление. Л., 1972.

Клюева Н. Н. Краткий словарь синонимов русского языка. М., 1956.

Кобрин Р. Ю., Пекарская Л. А. Лингвостатистический анализ употребления терминов нормативных словарей и ГОСТов в реальных научно-технических текстах. — В кн.: Языковая норма и статистика. М., 1977.

Кожина М. Н. О речевой системности научного стиля сравнительно с некоторыми другими. Пермь, 1972.

Козинцева Н. А. Видовые значения форм прошедшего времени в армянском языке. — В кн.: Проблемы лингвистической типологии и структуры языка. Л., 1977.

Кокрен У. Методы выборочного исследования. М., 1976.

Комиссаров В. Н. Словарь антонимов современного английского языка. М., 1964.

Кондаков Н. И. Логический словарь-справочник. Изд. 2-е. М., 1976.

Конецкая В. П. Характеристика лексических омонимов — слов, генетически связанных, и пути их образования в английском языке. — В кн.: Исследования по английской лексикологии. М., 1961.

Костюченко Ю. П. Количество и место омонимов в современном английском языке. (По материалам словарей среднего объема — 60 000 слов). Автореф. канд. дис. Харьков, 1954.

Кудрявский Д. Н. К статистике глагольных форм в Лаврентьевской летописи. — Известия отделения русского языка и словесности АН, т. 14, кн. 2. СПб., 1909.

Кудрявский Д. Н. К истории русского прошедшего времени. — РФВ, т. 65, № 1, 1911.

Кудрявский Д. Н. Древнерусские причастия настоящего времени действительного залога на *-а*. — РФВ, т. 68, № 4, 1912.

Ланге О., Банасиньский А. Теория статистики. М., 1971.

Левковская К. А. Лексикология немецкого языка. М., 1956.

Лесскис Г. А. О зависимости между размером предложения и характером текста. — ВЯ, 1963, № 3.

Лесскис Г. А. Некоторые статистические характеристики простого и сложного предложения в русской научной и художественной прозе XVIII—XX вв. — РЯНШ, 1968, № 2.

Львов М. Р. Словарь антонимов русского языка. Около 2000 антонимических пар. М., 1978.

Малаховский Л. В. Об использовании частотного словаря в лексикологическом исследовании. (На материале английского языка). — В кн.: *ИСИАПЛ, 1968.*

Малаховский Л. В. Омонимы: терминология, классификация, определение и принципы выделения. — В кн.: Лингвистические исследования, 1972. Ч. 2. М., 1973.

Малаховский Л. В. Омонимическая группа как лексическая микросистема. — В кн.: Лингвистические исследования, 1975. Вопросы строя индоевропейских языков. Ч. 2. М., 1975.

Малаховский Л. В. Омонимия слов как абсолютная лингвистическая универсалия. — В кн.: Иностранные языки в высшей школе, вып. 12. М., 1977.

Мартин Дж. Организация баз данных в вычислительных системах. М., 1978.

Марчук Ю. Н. Опыт машинной реализации дистрибутивной методики определения лексических значений. — В кн.: СтРААТ, 1973.

Маслов Ю. С. Омонимы в словарях и омонимия в языке (к постановке вопроса). — В кн.: Вопросы теории и истории языка. Л., 1963.

Маслова-Лашанская С. С. Лексикология шведского языка. Л., 1973.

Машкина Л. Е. Способ определения репрезентативности выборок. — В кн.: СТ I, 1969.

Никонов В. А. Длина слова. — ВЯ, 1978, № 6.

Перебейнос В. И. Об установлении статистических параметров авторских стилей. — В кн.: ЧСиАПЛТ, 1968.

Пиотровский Р. Г., Чижаковский В. А. О двуязычной ситуации. — В кн.: СтРААТ, 1971.

Пиотровский Р. Г. Текст, машина, человек. Л., 1975.

Пиотровский Р. Г., Бектаев К. Б., Пиотровская А. А. Математическая лингвистика. М., 1977.

Пиотровский Р. Г. Инженерная лингвистика и лингвистические автоматы. — Вестник АН, № 19, М., 1978.

Пиотровский Р. Г. Инженерная лингвистика и теория языка. Л., 1979.

Русова Н. Ю. Опыт статистического анализа стилевой дифференциации структуры простого предложения. — В кн.: ВСС, 1974.

Савицкий Н. П. Об устойчивости относительных частот лингвистических элементов. — Československá rusistika, 1966, № 4.

Семенова С. Т. Опыт отбора наиболее употребительных слов английского языка на основе частотных данных. — ИЯШ, 1975, № 1.

Сиротинина О. Б. Порядок слов в русском языке. Саратов, 1965.

Скороходько Э. Ф. Лингвистические проблемы обработки текстов в автоматизированных информационно-поисковых системах. — В кн.: Вопросы информационной теории и практики, сб. 25. М., 1974.

Слепак Б. Я. Функционирование синтаксических единиц как параметр стиля в современном английском языке. Автореф. канд. дис. Киев, 1978.

Словарь современного русского литературного языка. Т. 1—17. М.—Л. 1948—1965.

Смирницкий А. И. Лексикология английского языка. М., 1956.

Смирницкий А. И. Морфология английского языка. М., 1959.

Ставровский Н. В. Лингвистическая статистика как наука. — РЯНШ, 1971, № 1.

Статистичні параметри стилів. Київ, 1967.

Строева Т. В. Сопоставительная статистика падежных словоформ имени существительного в немецком и русском языках. — ИЯШ, 1968, № 5.

Тоом А. Л. Несимметричная коммуникация, фокализация и управление в играх. — Семиотика и информатика, вып. 7. М., 1976.

Трибис Л. И. Об одной модели распознавания лексических значений неоднозначных слов. — В кн.: СтРААТ, 1973.

Троянская Е. С. Актуальные проблемы исследования функциональных стилей. — В кн.: Лингвистические исследования научной речи. М., 1979.

Тышлер И. С. К вопросу о судьбе омонимов (на материале современного английского языка). — ВЯ, 1960, № 5.

Тышлер И. С. Словарь омонимов современного английского языка. Саратов, 1963.

Тышлер И. С. О проблемах омонимии в английском языке. Автореф. канд. дис. М., 1966.

Тышлер И. С. Словарь лексических и лексико-грамматических омонимов современного английского языка. Саратов, 1975.

Урбутис В. Способы образования лексических омонимов в литовском языке. Автореф. канд. дис. Вильнюс, 1956.

Фрумкина Р. М. Объективные и субъективные оценки вероятностей слов. — ВЯ, 1966, № 2.

Фрумкина Р. М. Вероятность элементов текста и речевое поведение. М., 1971.

Фрумкина Р. М. Статистические методы и стратегия лингвистического исследования. — Известия АН СССР. Серия литературы и языка, т. 34, № 2. М., 1975.

Шайкевич А. Я. Источники лексической омонимии в германских языках. Канд. дис. Т. 1—2. М., 1962. (Хранится в Библиотеке им. В. И. Ленина в Москве).

Шингарева Е. А. Синтаксис, семантика и прагматика информационного языка объектно-признакового типа. Автореф. канд. дис. Л., 1979.

Шрейдер Ю. А. Равенство, сходство, порядок. М., 1971.

Щедровицкий Г. П. Смысл и значение. — В кн.: Проблемы семантики. М., 1974.

Эман Э. Об омонимии в немецком языке. — ВЯ, 1960, № 5.

Якубайтис Т. А., Стурите Б. А. О статистической однородности текстов. — В кн.: ВСС, 1974.

Allén S. Nusvensk frekvensordbok baserad på tidningstext. V. 1—2. Stockholm, 1970—1971.

Bailly R. Dictionnaire des synonymes de la langue française. Paris, 1967.

Bally Ch. Linguistique générale et linguistique française. 2-e éd. entièrement refondue. Berne, 1944. (Русский перевод: Балли Ш. Общая лингвистика и вопросы французского языка. М., 1955).

Barnhart C. L. (ed.) The world book dictionary. V. 1—2. Chicago, 1974.

Bridges R. S. On English homophones. — Society for Pure English, v. 1. Tract No. 11. Oxford, 1919.

Carroll J. B., Davies P. D., Richman B. The American Heritage word frequency book. New York, 1971.

Cramer P. A study of homographs. — In: Norms of word association. New York — London, 1970.

De Luca A., Termini S. A definition of non-probabilistic entropy in the setting of fuzzy sets theory. — *IC, v. 20, No. 6, 1972.*

Dernoschek O. De elegantia Caesaris sive de commentariorum de B. G. et de B. C. differentiis animadversiones. Lipsiae, 1903.

Dreyfus H. L. What computers can't do. A critique of artificial reason. New York, 1972. (Цит. по русскому переводу: Дрейфус Х. Чего не могут вычислительные машины. Критика искусственного разума. М., 1978).

Dupuis H. Dictionnaire des synonymes et des antonymes. Paris, 1961.
Entwisle D. R. Form class and children's word associations. — JVL, v. 5, 1966.
Fourquet J. Grammaire de l'allemand. Paris, 1952.
Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. — JASA, v. 32, No. 200, 1937.
Gaines B. R., Kohout L. J. The fuzzy decade: a bibliography of fuzzy systems and closely related topics. — IJMMS, v. 9, No. 1, 1977.
Ganshina M. A., Vasilevskaya N. M. English grammar. 8th ed. Moscow, 1958.
Gilliéron J. Pathologie et thérapeutique verbale. Paris, 1921.
Grand Larousse de la langue française. Paris, 1971—1972.
Guiraud P. Problèmes et méthodes de la statistique linguistique. Paris, 1960.
Hall J. F. Learning as a function of word frequency. — AJP, v. 67, No. 1, 1954.
Herdan G. The advanced theory of language as choice and chance. Berlin—Heidelberg — New York, 1966.
Horn E. A basic writing vocabulary. 10 000 words most commonly used in writing. Iowa City, 1926.
Hornby A. S., Gatenby E. V., Wakefield H. The advanced learner's dictionary of current English. 2nd ed. London, 1958.
Hornby A. S., Gatenby E. V., Wakefield H. The advanced learner's dictionary of current English. Second edition. London, 1972.
Howes D. Application of the word frequency concept to aphasia. — In: Disorders of Language. London, 1964.
Howes D. H. and Solomon R. L. Visual duration threshold as a function of word probability. — JEP, v. 41, No. 6, 1951.
Jespersen O. Monosyllabism in English. — Proceedings of the British Academy, v. 14. London, 1928.
Jespersen O. A modern English grammar on historical principles. London, 1933.
Kieft A. Homonymie en haar invloed op de taalontwikkeling. Gröningen, 1938.
Leah L. L. Homonyms and synonyms as retrieval cues. — JEP, v. 96, No. 2, 1972.
Liebich B. Kleine Beiträge zur deutschen Wortforschung. — Beiträge zur Geschichte der deutschen Sprache und Literatur, Bd 29. Halle (Saale), 1898.
Mandelbrot B. Structure formelle des textes et communication. — Word, v. 10, No. 1. New York, 1954.
Menner R. J. The conflict of homonyms in English. — Language, v. 12, No. 4. Baltimore, 1936.
Miller G. A. Communication. — Annual Review of Psychology, v. 5. Palo Alto, Calif., 1954.
The new American Roget's college thesaurus in dictionary form. New York and Toronto, London, 1962.
Öhmann E. Über Homonymie und Homonyme im Deutschen. — Annales Academiae Scientiarum Fennicae, series B, t. 32, No. 1. Helsinki, 1934.
Pei M. The story of language. 2nd ed. London, 1966.
Peters H. N. The relationship between familiarity of words and their memory value. — AJP, v. 48, 1936.
Philipon-la-Madelaine, L. Des homonymes français, ou Mots qui dans notre langue se ressemblent par le son et diffèrent par le sens. Nouvelle éd. Paris, 1802.
Phoneme—grapheme correspondences as cues to spelling improvement. Washington, 1966.
Rath R. Trennbare Verben und Ausklammung. — Wirkendes Wort, Jg. 15, H. 4. Düsseldorf, 1965.
Rinsland H. D. A basic vocabulary of elementary school children. New York, 1945.
Shelling T. The strategy of conflict. Cambridge, 1960.
Skeat W. W. An etymological dictionary of the English language. Oxford, 1909.
Sumby H. Word frequency and serial position effect. — JVL, v. 1, 1963.
Thorndike E. L. The teacher's word book. New York, 1921.
Thorndike E. L. and Lorge I. The teacher's word book of 30 000 words. New York, 1944.
Trésor de la langue française. Dictionnaire de la langue du XIX^e et du XX^e siècle (1789—1960) publié sous la direction de Paul Imbs. T. 1 f. Paris, 1971—75.
Ullmann S. Language and style. Oxford, 1964.
Ulvestad B. Statistik und Sprachbeschreibung. — In: Das Ringen um eine neue deutsche Grammatik. Darmstadt, 1962.
Webster's third new international dictionary of the English language. Unabridged. Springfield, Mass., 1961, 1966.
Webster's dictionary of synonyms. Springfield, Mass., 1968.
Webster's seventh new collegiate dictionary. Springfield, Mass., 1971.
Williams E. R. The conflict of homonyms in English. — Yale Studies in English, v. 100. New Haven — London, 1944.
Yule G. U. On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. — Biometrica, v. 30, No. 3—4. 1939.
Zadeh L. A. Outline of a new approach to the analysis of complex systems and decision processes. — IEEE TSMC, v. SMS, 3 Jan., 1973. (Цит. по русскому переводу: Заде Л. А. Основы нового подхода к анализу сложных систем и процессов принятия решений. — Математика сегодня, 7, 1974).
Zadeh L. A. The concept of a linguistic variable and its application to approximate reasoning. New York, 1973. (Цит. по русскому переводу: Заде Л. А. Понятие лингвистической переменной и его применение к принятию приближенных решений. М., 1976).
Zipf G. K. The psycho-biology of language. Boston, 1935.

СОКРАЩЕНИЯ И УСЛОВНЫЕ ОБОЗНАЧЕНИЯ

АД — агрегат данных.

АПТ — автоматическая переработка текста; сб.: Автоматическая переработка текста. Кишинев, 1978.

АРСл — Англо-русский словарь. Сост. В. К. Мюллер. Изд. 13-е. М., 1971.

АС — автоматический словарь.

АСл — автоматический словник.

БАРС — Большой англо-русский словарь. Т. 1—2. М., 1972.

БД — банк данных.

БОЛИД — база основных лингвистических данных.

ВСС — сб.: Вопросы статистической стилистики. Киев, 1974.

ВЯ — журн.: Вопросы языкознания, Москва.

ГЛ — гиперлексема.

ДСП — дифференциальный семантический признак.

ЕЯ — естественный язык.

ИБ — информационная база.

ИЛ — инженерная лингвистика.

ИО — информационное обеспечение.

ИР — искусственный разум.

ИП — информационный процесс.

ИЯ — искусственный язык.

ИЯШ — журн.: Иностранные языки в школе, Москва.

КК — коэффициент комплексности.

КП — ключ поиска.

ЛА — лингвистический автомат.

ЛЕ — лексическая единица.

ЛЗ — лингвистическая дача.

ЛО — лингвистическое обеспечение.

МАРС — многоцелевой автоматический русский словарь.

МО — машинная основа.

МП — перевод с помощью ЭВМ.

НТИ — журн.: Научно-техническая информация, Москва.

ОГ — омонимическая группа, омогруппа.

ОК — омокомплект.

ОНСС — омонимическая насыщенность словарного состава.

ОП — основная память.

ОС — операционная система.

ОСл — обратный словарь русского языка. М., 1971.

ПО — предметная область.

ПП — поле признаков.

ППр — поисковая процедура

РСФ — репрезентативная словоформа.

РФВ — журн.: Русский филологический вестник, Варшава.

РЭН — распознающий эталонный набор.

РЯНШ — журн.: Русский язык в национальной школе, Москва.

РЯШ — журн.: Русский язык в школе, Москва.

СемП — семантический признак.

СК — сжатый код.

СлС — Словарь синонимов. Справочное пособие. Л., 1975.

СП — семантическое пространство.

СС — семантическая сеть.

с/с — словосочетание.

ССт — словарная статья.

СТІ — сб.: Статистика текста. Сборник статей. Материалы семинара «Общие проблемы языкознания и лингвостатистические методы исследования» (1968—1969 гг.). Т. 1. Лингвостатистические исследования. Минск, 1969.

СтР — группа «Статистика речи»; сб.: Статистика речи. Л., 1968.

СтРААТ — сб.: Статистика речи и автоматический анализ текста. Л., 1971, 1973, 1974; М.—Л., 1979.

с/у — словоупотребление.

СУБД — система управления базой данных.

СУВВ — система управления вводом-выводом.

с/ф — словоформа.

ТИСО — тип интеркомпонентных смысловых отношений.

ФРС — формальное распознавание смысла.

ЦП — цельное предложение.

ЧВААТ — сб.: Частные вопросы автоматического анализа текста. Минск, 1972.

ЧСиАПЛТ — сб.: Частотные словари и автоматическая переработка лингвистических текстов. Минск, 1968.

ЭД — элемент данных.

ЭП — элементарное предложение.

ЭПр — элементное пространство.

a, A — adjectif, adjective.

adv. — adverbe, adverb.

AJP — American Journal of Psychology, New York.

art — article.

c, cj — conjonction, conjunction.

COD — Concise Oxford dictionary of current English. Ed. by H. W. Fowler and F. G. Fowler. 4th ed. London, 1964.

dem — demonstrative.

F — абсолютная частота.

F* — абсолютная накопленная частота.

ger — gerund.

i, int — interjection.

IC — журн: Information and Control, New York — London.

IEEE — Institute of Electrical and Electronics Engineers.

IEEE TSMC — журн.: IEEE Transactions on Systems, Man and Cybernetics, New York.

IJMMS — International Journal of Man-Machine Studies, New York.

imp	— imperative.	n, N	— noun.
inf	— infinitive.	num	— numeral.
JASA	— Journal of the American Statistical Association, New York.	p	— person.
JEP	— Journal of Experimental Psychology, Washington.	pl	— plural.
JVL	— Journal of Verbal Learning and Verbal Behavior, New York.	poss	— possessive.
m, M	— количество лингвистических единиц, обладающих данной частотой.	pp	— past participle.
N	— объем выборки.	prep, Prep	— preposition.
		pres	— present.
		pron	— pronom, pronoun.
		pt	— past.
		s	— substantif, substantive.
		sbj, subj	— subiunctive (mod).
		sg	— singular.
		v, V	— verb, verbe.

СОДЕРЖАНИЕ

СТАТИСТИКА РЕЧИ
АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТА

Утверждено к печати
Институтом языкознания АН СССР

Редактор издательства *Ю. Ф. Денисенко*
Художник *М. И. Разумевич*
Технический редактор *А. П. Чистякова*
Корректоры *А. И. Кац* и *С. Н. Семиглазова*

ИБ № 9032

Сдано в набор 25.02.80. Подписано к печати 24.07.80. М-38772. Формат 60 × 90¹/₁₆. Бумага типографская № 1. Гарнитура обыкновенная. Печать высокая. Печ. л. 14+1 вкл. (⁶/₈ п. л.)=14.62 усл. печ. л. Уч.-изд. л. 15.14. Тираж 1750. Изд. № 7483. Тип. зак. 1193. Цена 2 р. 10 к.

Издательство «Наука» Ленинградское отделение
199164, Ленинград, В-164, Менделеевская лин., 1

Ордена Трудового Красного Знамени
Первая типография издательства «Наука»
199034, Ленинград, В-34, 9 линия, 12