Юхан Тулдава

ПРОБЛЕМЫ И МЕТОДЫ КВАНТИТАТИВНО--СИСТЕМНОГО ИССЛЕДОВАНИЯ ЛЕКСИКИ

ТАРТУСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Юхан Тулдава

ПРОБЛЕМЫ И МЕТОДЫ КВАНТИТАТИВНО-СИСТЕМНОГО ИССЛЕДОВАНИЯ ЛЕКСИКИ

ТАРТУ 1987

Ответственный редактор доктор филологических наук
Р.Г.Пиотровский

Рецензенты доктор филологических наук В.И.Перебейнос
и доктор филологических наук Т.-Р.Вийтсо
Тулдава Ю.А.

Проблемы и методы квантитативно-системного исследования лексики. – Таллин: Валгус, 1987. – 204 с.

Цель книги – обобщить опыт квантитативно-лингвистических исследований на современном этапе развития, разработать целостную концепцию квантитативно-системного подхода к исследованию лексики и реализовать ее на иллюстративном материале из разных языков.

Работа адресована специалистам в области квантитативной лингвистики, а также другим языковедам, интересующимся расширением арсенала методов лингвистического исследования. Книга может быть использована в качестве учебного пособия на отделениях структурной и математической лингвистики вузов.

T 4602000000-329
902(15) –87 заказное 8I.2

Тартуский государственный университет. Юхан Т у л д а в а. Проблемы и методы квантитативно-системного исследования лексики. На русском языке. Художник-оформитель Т.Ару. Таллин, "Валгус". Редактор Р.Г.Пиотровский. Художественный редактор Р.Эйлсен. Подписано в печать I0.07.87. МВ- 04389. Формат 60x90/I6. Бумага офсетная. Машинопись. Ротапринт. Усл. печ. л. I2,75. Усл. кр.-отт. I3,0. Уч.-изд.л. I2,38. Тираж 600 экз. Заказ № 682. Цена I.90. Заказное. Издательство "Валгус", 200090 Таллин, Пярнуское шоссе, I0. Типография ТГУ, 202400 Тарту, ул. Тийги, 78.

ПРЕДИСЛОВИЕ

В настоящей работе осуществляется квантитативно-системный подход к изучению лексики, т.е. разрабатывается комплекс исследовательских приемов и научных объяснений при систематическом исследовании лексики языка в квантитативном освещении. Реализация этой задачи концентрируется вокруг трех основных проблем:

- теоретико-методологические основы и методика исследования;
- квантитативные закономерности организации лексики (в словаре и тексте);
- конкретный квантитативно-лингвистический анализ лексики и интерпретация результатов исследования (иллюстративный материал взят из разных языков).

В работе дается обобщенное и целостное представление о концептуальном аппарате предлагаемого квантитативно-системного подхода. При этом уточняются и заново определяются такие научные понятия, как квантификация, измерение, распределение, вероятностная система и др. применительно к материалу и проблематике квантитативно-системного исследования лексики в свете требований философской и общенаучной методологии (диалектическая связь необходимости и случайности, количества и качества; всеобщая связь, системность). По-новому трактуются проблемы единиц и уровней исследования, типологии лингвистических распределений, стратификации лексики и классификации лексических групп. Предлагается модель речевой деятельности на основе перекрещивающихся факторов "потенции - реализации" (языка - речи) и "статики - динамики". Эта модель позволяет четко разграничить предмет и сформулировать задачи квантитативно-лингвистического исследования. Особое внимание уделяется историко-социальным характеристикам лексики с квантитативной точки зрения. В интерпретационном плане дается принципиально новое решение многим традиционным проблемам квантитативной лингвистики (рост и развитие лексики, зависимость объемов словаря и текста, лексическая связь

текстов, лексическое богатство и др.). В квантитативном плане исследуются некоторые аспекты семантического и стилистического анализа лексики, выявляются внутрисистемные и внесистемные связи лексических систем с помощью новых вариантов методов "объединения словарей" и кластерного анализа.

Настоящая работа представляет собой сокращенный вариант исследования автора о квантитативно-системном анализе лексики эстонского языка (Тулдава Ю.А., 1984) с привлечением новых данных по разным другим языкам. Основной целью книги является обобщение опыта квантитативно-лингвистических исследований на данном этапе развития и ознакомление читателей с некоторыми новыми теоретическими установками и практическими методами в современной квантитативной лингвистике.

I. ТЕОРЕТИКО-МЕТОДОЛОГИЧЕСКИЕ ОСНОВЫ КВАНТИТАТИВНО-СИСТЕМНОГО АНАЛИЗА ЛЕКСИКИ

В условиях научно-технической революции и информационного взрыва, характерного для нашего времени, проблемы теории и методологии науки приобретают первостепенную значимость. В настоящее время квантитативная лингвистика достигла такого уровня, что необходимо обобщать теоретические выводы, уточнить понятия и методы исследования, а также раскрыть некоторые новые стороны объекта и наметить пути их объяснения. В частности, исследование лексики с системных позиций и с применением квантитативных методов требует разработки методологических основ комплексного квантитативно-системного подхода. Это позволяет по-новому осмыслить некоторые основные проблемы квантитативной лингвистики с единой точки зрения.

I.I. ОБОСНОВАНИЕ КВАНТИТАТИВНО-СИСТЕМНОГО ПОДХОДА К ИЗУЧЕНИЮ ЛЕКСИКИ

Предмет и исходные принципы исследования. Если объектом лексикологического исследования следует считать строение и функционирование лексики данного языка (или группы языков), то ввиду сложности, многогранности и многоплановости описываемого объекта при конкретном подходе требуется ограничить область изучения и уточнить аспект, или сторону объекта, подвергаемую рассмотрению. В методологии науки такой ограниченный в определенном смысле объект называется предметом данного исследования. Это означает, что познание объекта как онтологической сущности представляет собой многосторонний процесс, в котором выделяются различные гносеологические аспекты. В строении и функционировании лексики как объекта исследования можно вычленять разные стороны ее проявления, которые будут представлять собой различные предметы исследования, изучаемые с помощью специфических методов. Существуют, например, историко-сопоставительный, структуралистский, лингвостатистический и др. подходы к изучению лексики. Каждый раз наличный предмет ставится уже в исходной точке его ана-

лиза в контекст нового объяснения, а это отвечает принципу множественности описаний любого объекта реального мира и, в конечном счете, отражает диалектическое соотношение многообразия мира и единства мира. Для охватывания объекта в целом обязательно требуется совмещение различных точек зрения, их синтез, который, однако, имеет своей предпосылкой углубленный анализ отдельных сторон объекта.

Выделение предмета исследования определяется как поставленной задачей, так и объемом и глубиной знаний в данной области, которыми уже располагает наука. Исходя из общей задачи описания квантитативных сторон словарного состава конкретного языка в связи с выяснением некоторых важных закономерностей строения и функционирования лексики в речевой деятельности вообще, мы должны в то же время учитывать уровень развития квантитативной лингвистики в настоящее время и возможности синтеза имеющихся знаний в стройную систему. Мы принимаем определение квантитативной лингвистики, данное Р.Г. Пиотровским, К.Б. Бектаевым и А.А. Пиотровской (1977, с. 8), которые под этим понятием подразумевают изучение и экспликацию лингвистических явлений с помощью методов "количественной" математики (теория вероятностей, математическая статистика, теория информации и др.).+ Квантитативная лингвистика противопоставляется "комбинаторной" лингвистике, опирающейся на разделы "неколичественной" математики (теория множеств, математическая логика, теория алгоритмов и т.д.). Квантитативная и комбинаторная лингвистики являются двумя сторонами общего, родового понятия "математическая лингвистика". Особый раздел в математической лингвистике составляет т.н. статистико-комбинаторное моделирование (Андреев Н.Д., 1967), в котором сочетаются теоретико-множественный, алгоритмический и статистический подходы к исследованию языка.

Практические и теоретические исследования в области математической лингвистики за последнее десятилетие продвинулись вперед в такой степени, что созрела почва для новых синтезирующих подходов при изучении квантитативного аспекта лексики. Выяснились два основных принципа таких исследований: п р и н ц и п с и с т е м н о с т и и в е р о я т н о с т н о - с т а т и с т и ч е с к и й х а р а к т е р

+ Частным случаем квантитативной лингвистики следует считать "лингвостатистику", т.е. изучение лингвистических объектов с помощью традиционных статистических методов.

организации лексики. Объединение этих принципов дает нам системно-вероятностный подход, который с успехом практикуется во многих работах последнего времени (Алексеев П.М., 1977; Бектаев К.Б., 1978; Нелюбин Л.Л., 1983; и др.). На новом этапе развития квантитативной лингвистики выяснилось, что можно еще более точно определить и более углубленно рассмотреть такие основные понятия, как вероятностная система в лингвистике и методы исследования закономерностей функционирования вероятностных систем. Оказалось, что кроме вероятностно-статистических и теоретико-информационных методов большую роль при описании и объяснении системных аспектов лексики могут играть и некоторые другие методы количественной математики, например, математический анализ, в том числе и теория функций (Пиотровский Р.Г. и др., 1977). В последнее время к исследованию систем и подсистем в области лингвистики привлекается и теория нечетких множеств (Заде Л., 1976; Пиотровский Р.Г., 1979). Характерной чертой новейших исследований по квантитативной лингвистике является стремление разработать теоретические основы квантитативной типологии текста в рамках общей теории лингвистики текста (например, Алексеев П.М., 1981).

Расширение арсенала исследовательских приемов квантитативной лингвистики за счет привлечения новых количественных методов математики в дополнение к традиционным вероятностно-статистическим методам вызывает необходимость уточнить название общего подхода к изучению квантитативных свойств лингвистических объектов. Представляется, что наиболее подходящим является обобщенное название – к в а н т и т а т и в н о – с и с т е м н ы й а н а л и з (Тулдава Ю.А., 1979), который указывает на системность объекта, изучаемого с помощью различных методов количественной математики. В то же время согласно нашей концепции остается в силе тезис о рассмотрении лексики (и других лингвистических объектов) в виде вероятностной системы, характеризующейся не только случайностью параметров (низший уровень организации), но и определенной устойчивостью и регулярностью в массе случайных событий (высший уровень организации). Таким образом, в вероятностных системах в современном понимании не исключается детерминированность связей, но она переносится на высший, обобщенный уровень организации (Сачков Ю.В., 1971). В философском осмыслении здесь просматривается диалектическая связь категорий случайности и необходимости, которая в соединении

с принципом всеобщей связи явлений объективной действительности (как основание системного подхода к исследованию предметов и явлений реального мира) представляют собой философско-методологические постулаты развиваемого в данной работе квантитативно-системного подхода к изучению лингвистических объектов.

Итак, п р е д м е т о м исследования при квантитативно-системном подходе (в вероятностной интерпретации) следует считать квантитативные свойства и закономерности строения и функционирования лексики (или других лингвистических объектов), рассматриваемые с системных позиций и с упором на вероятностную природу функционирования языка. Лексика рассматривается как вероятностная система со всеми свойствами, присущими таким системам, в том числе со свойствами устойчивости и вариативности. Подобно всем системам, лексика распадается на различные подсистемы и может рассматриваться как многоплановое и многоуровневое образование; в то же время она сама является подсистемой (элементом) общей системы языка (речевой деятельности) и составляет определенный уровень в иерархии языковых явлений.

Квантитативно-системный подход к изучению лексики, опирающийся на системность и вероятностно-квантитативный аспект функционирования языка, является центральным понятием и обобщающим принципом данного исследования. Этот основной принцип должен объединить все остальные понятия, суждения и законы в определенную целостность и дать общую направленность последующему исследованию эмпирических фактов. Прежде чем приступить к изучению фактического материала и основных закономерностей функционирования лексики, требуется более подробно рассмотреть и конкретизировать некоторые наиболее общие понятия и категории, раскрывающие и уточняющие упомянутый основной методологический принцип исследования. При этом нельзя забывать и о том, что объектом описания является определенный аспект речевой деятельности (см. гл. I.2) и в конце концов решается проблема, относящаяся к области лингвистики. Известно, что всякие принципы и категории могут служить научному изучению какого-либо объекта лишь тогда, когда они предварительно получают соответствующую переработку и интерпретацию, в данном случае в свете требований квантитативно-системного анализа лексики. Рассмотрим ниже в таком освещении понятие системности лексики, ее квантитативный и вероятностный аспекты, а также ряд производных понятий и положений,

входящих в концептуальный аппарат квантитативно-системного подхода к изучению лексики языка.

Системность языка и лексики. Большинство лингвистов в наше время сходится на том, что системность является одной из наиболее существенных объективных черт языка. "Странно было бы отрицать наличие системности в языке, - отмечает Б.А. Серебренников, - поскольку системность языка обусловлена не только основной функцией языка быть средством общения, но и необходимо связана с некоторыми чисто физиологическими и психологическими особенностями человека, не говоря уже о том, что основанием ее может служить наличие системности в окружающем человека мире". (Общее языкознание, 1973, с. 295). Методологические проблемы системного подхода на конкретных примерах развития и функционирования языка не раз обсуждались нашими языковедами. Отмечалось, что в принципе "можно считать установленным, что язык относится к системным образованиям. Однако термины "система" и "системный" в разных работах понимаются по-разному " (Солнцев В.М., 1977, с. 12).

Вместе с тем многие авторы указывают на то, что системный характер л е к с и к и (словарного состава языка) в рамках общей системы языка имеет свои особенности. Это связано с тем, что лексика представляет собой особо сложный и противоречивый объект, нелегко поддающийся строгой систематизации и классификации. Системность лексического уровня изучена гораздо меньше по сравнению с некоторыми другими уровнями.

Трудности системного анализа лексики объясняются ее "сущностными характеристиками", такими как неисчисляемость ее единиц, неограниченные возможности комбинаторики слов, сложность и неоднородность типов словесной связи в системе языка и в речи, экстралингвистическая детерминированность слов и т.д. Все это создает дополнительные трудности при изучении лексики, которая порой предстает как "плохо организованная, диффузная система", но которая тем не менее успешно функционирует в процессе общения людей. Создается впечатление, что лексика не нуждается в строго регулярной и жесткой организации, чтобы функционировать, так как именно эластичность лексической системы придает гибкость и маневренность речевой деятельности.

Суждения о системности лексики (как и языка в целом) зависят, конечно, от того, как определять поня-

тие "система". Когда некоторые исследователи говорят о явлениях "антисистемности" в языке (Будагов Р.А., 1978; Филин Ф.П., 1979) или о диалектическом единстве и взаимопроникновении системных и асистемных процессов (Маковский М.М., 1980), то, очевидно, под системой понимается жесткая регулярность отношений и связей элементов, стабильность и непротиворечивость взаимодействия компонентов целого. Это, собственно, признаки т. наз. динамических систем, которые существуют наряду с вероятностными (статистическими) системами, имеющими несколько иные свойства. В данной работе речь идет о системном подходе в современном понимании, когда возможны различные конкретные пути - как детерминированные, так и вероятностные - к решению проблем.

Проявление системного подхода в наши дни связано в первую очередь с необходимостью исследования больших сложных систем, которые, как правило, слабо структурированы и которые содержат частично неформализуемые элементы, причем функционирование таких систем часто происходит в условиях неопределенности. Именно речевая деятельность в целом, а также ее подсистемы, в том числе лексика, относятся к таким сложным образованиям, познание которых настоятельно требует системного подхода.

В наиболее общей форме с и с т е м у можно определить как "целостный объект, состоящий из элементов, находящихся во взаимных отношениях" (Солнцев В.М., 1977, с. 14). Совокупность отношений между элементами называется с т р у к т у р о й данной системы, причем в определениях системы обычно подчеркивается относительно устойчивый характер отношений между элементами, что является необходимой предпосылкой стабильности системы и успешного ее функционирования. "Структура, - пишет Н.Ф. Овчинников (1966, с. 268), - обеспечивает сохранение системы, и потому она есть то, что остается относительно постоянным, сохраняющимся в изменчивом бытии вещи."

Таким образом, структура системы, или сеть устойчивых отношений между элементами, представляет собой важнейший аспект системы, и задача исследователя состоит в том, чтобы найти адекватный способ ее описания. Однако следует заметить, что хотя структура в определенных условиях может быть объектом изучения в отвлечении от субстанции данной системы, но, по существу, структура не может существовать вне с у б с т а н ц и и (э л е м е н т о в) системы, т.е. структуру и субстанцию следует рассматривать в их диалектической вза-

имосвязи.

Материальный состав (субстанция, элементы) и соотношение элементов (структура) составляют вместе взятые "строение" целостной системы. Однако системный подход требует совместного рассмотрения строения и функционирования системного объекта. Под функционированием системы понимается – в широком смысле – взаимодействие системы с внешней средой, в том числе с другими системами и подсистемами. В более узком плане функционирование системы – это взаимодействие ее элементов в рамках сложившейся структуры. Очевидно, что функционирование составляет такой ингредиент системы, который неразрывно связан со строением системы. Только синтез, диалектическое единство трех главных ингредиентов системы – множества элементов, структуры и функционирования обеспечивает существование целостной системы. При этом структуру и функционирование в их единстве можно рассматривать как "организацию" системы. В таком случае система предстает как множество элементов плюс организация (структурная и функциональная).

Само понятие лексика понимается в двух основных значениях. С одной стороны, лексика рассматривается как совокупность отдельных единиц (словоформ или лексем). В таком случае исследуются системные связи и закономерности словаря и текста в направлении от составных частей к целостности. С другой стороны, лексика рассматривается как единая совокупность, в которой выявляются различные параметры (законы взаимозависимостей составных частей и т.д.), относящиеся к объекту в целом. В этом случае исследование идет от целостности к составным частям. Эти два основных направления (аспекта) системного исследования лексики можно назвать аналитическим и интегральным соответственно (ср. Общее языкознание, 1972, с. 78).

Системный характер лексики как единой совокупности обнаруживается, в первую очередь, в распределении слов по разного рода лексическим группам – подсистемам, в которых существуют определенные отношения и связи между элементами. Можно сказать, что лексика – это, прежде всего, совокупность различных подсистем. Внутренняя структура и условия функционирования этих подсистем еще недостаточно познаны из-за многочисленности и разнородности элементов и сложности отношений между ними. Нет специальных исследований, которые показывали бы, каким образом

эти подсистемы между собой системно связаны. Обычно только отмечается, что "отдельные звенья системы взаимосвязаны и взаимозависимы, но не строго уравновешены" (Общее языкознание, 1972, с. 52). В целом для лексической системы (подсистемы в рамках общей системы языка) характерно относительное или "подвижное" равновесие. При этом можно констатировать, что это есть "система, складывавшаяся и изменявшаяся стихийно на протяжении тысячелетий. Поэтому в каждом языке немало "нелогичного", "нерационального" и противоречивого " (Маслов Ю.С., 1975, с. 33).

Встает вопрос, какого рода должна быть та модель системы, которая адекватным образом отражала бы в едином комплексе все многообразие целостного объекта, включая моменты регулярности и нерегулярности (стабильности и нестабильности и т.д.), присущие языку в целом и его подсистеме – лексике в частности.

Вероятностная система. Очевидно, что если под системой понимается жесткая регулярность отношений и связей элементов, строго функциональная связь между компонентами целостного образования и т.п., то мы имеем дело только с одной из разновидностей системы, а именно с так называемой динамической системой. Но в современных системных исследованиях наряду с динамическими системами рассматриваются и в е р о я т н о с т н ы е системы, которые характеризуются тем, что в них целостность и устойчивость системы сочетается с достаточно широкой автономностью частей (Сачков Ю.В., 1971; Кравец А.С., 1976). В отношении таких систем можно говорить о регулярном чередовании "жестко детерминированного и вероятностно-статистического способов управления" (Блауберг И.В., 1977, с. 10). В исследованиях по теории вероятностных систем указывается, что параметры таких систем относятся к различным уровням, они как бы делятся на два класса: случайные события (на более низком уровне) и закономерности, регулярности в массе случайных событий (на более высоком уровне). При этом характеристики высшего уровня, определяя общую структуру системы, не определяют каждое конкретное случайное событие в отдельности, вернее, они определяют эти случайные события (как характеристики низшего уровня) лишь обобщенно, интегрально. При таком подходе становится ясным, что всякие "антисистемные" явления органически включаются в ткань системы как целостного образования, обладающего как регулярными, так и нерегулярными (случайными, переходными, вариатив-

ными и т.д.) свойствами.

Иногда говорят о "разной степени системности" языковых явлений (Общее языкознание, 1972, с. 74). Очевидно, здесь имеется в виду разная степень регулярности, симметричности и т.п. на различных уровнях языковой системы. В этом смысле можно говорить о "ядре" и "периферии" системы, относя к последней все нерегулярности и кажущиеся антисистемные явления. Действительно, в структуре и функционировании языка можно выделять как вполне регулярные, однозначно детерминированные, так и совершенно случайные или нерегулярные явления и процессы. Они образуют как бы противоположные полюсы (предельные точки) системы, в то время как большинство реальных языковых и речевых явлений занимает "промежуточное" положение. Это можно понимать как конкретизацию основного принципа вероятностных систем в том смысле, что параметры низшего уровня, являющиеся случайными событиями, в разной степени управляются параметрами высшего уровня детерминации. Но следует иметь в виду, что полная детерминированность сложной системы (или частей такой системы) – это скорее исключение, чем правило. Полная детерминированность ("динамическая закономерность") – по сути дела статистическая закономерность с вероятностью осуществления, равной I. Это создает возможность при последовательном вероятностном подходе о б о б щ а т ь в е р о я т н о с т н у ю м о д е л ь в т о м с м ы с л е, ч т о о н а с о д е р ж и т в с е б е д е т е р м и н и р о в а н н о с т ь к а к ч а с т н ы й с л у ч а й.

Важно отметить, что вероятностная трактовка явлений действительности связана не только с представлением о вероятностном характере знания, но и главным образом с представлением о том, что "сам объект познания в своем движении и изменении, в своих взаимоотношениях с другими объектами подчиняется вероятностным закономерностям" (Штофф В.А., 1972, с. 131). Следовательно, вероятностный подход к изучению системы обусловлен самим объектом исследования – сложным и многообразным явлением, в котором необходимая связь отдельных компонентов, их причинная обусловленность должны, очевидно, представляться не в виде чистого детерминизма, а чаще всего как диалектическая связь случайностей и скрывающейся за ними необходимости. Эта характеристика в полной мере относится и к объекту нашего исследования – лексике языка.

Вероятностный подход к исследованию систем имеет фило-

софское обоснование. Оно покоится на концепции о внутренней связи категорий необходимости и случайности, а также категорий сущности и явления (см. Сачков Ю.В., 1971, с.167 и след). На этой философской основе вырастает и получает в наши дни все большее распространение "вероятностный стиль мышления", который смело включает случайность как форму проявления необходимости в саму структуру теоретических систем. Это дает возможность перейти к исследованию весьма сложных объектов, характеризующихся взаимопроникновением в своей структуре жесткого и подвижного начал.

Следует отметить, что понятие в е р о я т н о с т и само по себе не является чисто математическим понятием. Вероятность может выражаться не только через статистические характеристики объектов (например, через относительную частоту), а через самые различные свойства объектов (полезность, ценность и т.д.). Обобщая, можно сказать, что понятие вероятности характеризует "любые процессы или ситуации, в которых объективно возможны альтернативные исходы" (Суслов И.П., 1978). При квантитативном подходе наибольший интерес представляет численное выражение степени вероятности, а это приводит к логико-математическому понятию вероятности, которая по определению является "объективной характеристикой степени возможности появления определенного события в каких-то заранее заданных условиях, которые могут повторяться неограниченное число раз" (Философская энциклопедия, т. I, с. 244). Это означает, что понятие вероятности применяется к массовым явлениям, случившимся очень много раз. А м а с с о в о с т ь употребления лингвистических единиц обеспечивается как раз функцией языка как средства коммуникации. При вероятностном подходе фактор массовости обычно сочетается с понятием с л у ч а й н о с т и. Это понятие используется для определения специфики массового явления и характеризует, таким образом, некоторое объективное положение вещей. Случайность возникает "в результате взаимодействия фактов или пересечения необходимостей" (Рузавин Г.И., 1978, с. 227). Известно, что слова употребляются в тексте не только в силу чистой случайности, но и в результате направленного выбора. Но в массе лингвистических событий (например, при порождении текста) направленный выбор подчиняется влиянию большого количества самых разнообразных, в том числе экстралингвистических, факторов, так что практически можно исходить из признания случайности появления лингвистических единиц в речевом

потоке. Однако поток индивидуальных "случайных" событий представляет собой лишь внешнее проявление внутренних, т.е. "необходимых" тенденций. Это отражает сущность вероятностных систем, в которых случайность и необходимость диалектически взаимосвязаны.

Момент н е о б х о д и м о с т и проявляется в вероятностной системе двояко: во-первых, как относительная стабильность частот отдельных элементов или групп элементов (это и есть внутреннее свойство вероятности, которая в реальности выступает как "тенденция частот при определенных условиях группироваться вокруг некоторого постоянного значения", см. Колмогоров А.Н., 1956, с. 274)+, во-вторых, в виде устойчивого распределения элементов, выражающего наличие внутренней упорядоченности в системе. Распределение как обобщенное, интегральное понятие является важнейшей структурной характеристикой вероятностной системы. В дальнейшем изложении будет показано, что устойчивые распределения присущи и предмету нашего исследования - квантитативному аспекту системы лексики.

Представление предмета исследования в виде вероятностной системы требует более подробного обсуждения условий и возможностей квантитативного подхода к изучаемым явлениям, в данном случае к явлениям лексики, а также выяснения существенного вопроса взаимосвязи между количественным и качественным аспектами языка.

Квантитативный аспект. В наше время нет необходимости защищать применение квантитативных критериев в лингвистических, в том числе лексикологических исследованиях. Квантитативные (особенно статистические) методы уже давно и прочно завоевали признание в языкознании. "Одним из реальных оснований применения квантитативных методов в изучении языка и речи нужно признать объективную присущность языку количественных признаков, количественных характеристик" - отмечает Б.Н. Головин в своей книге "Язык и статистика" (Головин Б.Н., 1971, с. 11). Для использования квантитативных методов в исследовании тех или иных объектов достаточно, "чтобы эти объекты демонстрировали повторяемость, периодичность своих свойств, обладали в той или иной степени инвариантными отношениями, имели закономерное распределение своих параметров и т.д." (Садовский В.Н., 1974, с. 42-43). Именно п о в т о р я е м о с т ь (рекуррентность, периодичность) языковых, в

+ О т.н. мизесовском подходе к математическому оформлению теории вероятностей см. Алимов Ю.И., 1980.

том числе лексических единиц, их воспроизведение в различных текстах является наиболее важным условием квантификации языкового материала и применения различных математических методов для его анализа.

Полезность и важность применения квантитативного подхода к изучению лингвистических объектов подчеркивали многие выдающиеся лингвисты как прошлого, так и настоящего времени. "Нужно чаще применять в языкознании количественное, математическое мышление," - указывал И.А. Бодуэн де Куртене (1963, с. 17). Академик В.В. Виноградов, говоря о частотности употребления разных типов слов в разных стилях книжной и разговорной речи, отмечал, что "точные изыскания в этой области помогли бы установить структурно-грамматические, а отчасти и семантические различия между стилями" (Виноградов В.В.,1938, с. 176-177). "Частотность принадлежит функциональной стороне языковой системы /.../ Учет частотности любого языкового явления - полезный прием при анализе," - пишет В.Н. Ярцева (1970). Вместе с тем некоторые исследователи, не отрицая важности применения математических методов в языкознании, указывают на их известную ограниченность. "Математические методы анализа языковых явлений, - пишет Ф.П. Филин, - /.../ которые обещают многое в развитии нашей науки, имеют свои пределы." По мнению Ф.П. Филина, "язык имеет не только и не столько количественную сторону, поддающуюся исчислению. Он существенно отличается от машинных и алгоритмических языков /.../ тем, что его элементы (слова, предложения, грамматические формы и пр.) многозначны, имеют свойства образовывать новые переносные значения и оттенки значений, не говоря уже о бесконечном разнообразии их употреблений /.../ Переплетение ассоциаций элементов языка настолько сложно и бесконечно (как само наше сознание), что не может поддаться самому изощренному количественному учету." (Филин Ф.П., 1979, с. 27).

Следует признать справедливость этого высказывания в том смысле, что математические методы - в данном случае имеются в виду квантитативные методы - действительно не в состоянии решать любые проблемы анализа языковых явлений. Квантитативный подход способен охватить лишь определенный аспект языка и речи. Но это - существенный аспект языка, отражающий ряд важных сторон речевой деятельности, которые невозможно обнаружить чисто качественным анализом. При квантитативном анализе приходится иногда упрощать лингвистическую действительность (например, когда при статистических подсчетах не

учитываются поливалентность и многообразие оттенков значений слов). Но при этом анализе в принципе возможен и более дифференцированный подход к полисемичности, поливалентности и другим свойствам слов – в такой мере, в какой эти дифференцированные данные может в четкой форме представить качественный анализ. Часто оказывается, однако, что "переплетение ассоциаций элементов языка настолько сложно и бесконечно", что оно не может поддаться в полной мере не только количественному, но и качественному анализу. Кроме того, следует указать еще на то, что чисто качественный анализ сложных явлений нередко остается на уровне субъективных, произвольных интерпретаций.

Как известно, к а ч е с т в о и к о л и ч е с т в о – это парные категории диалектики и через них "вещь осмысливается полностью, без остатка" (Шептулин А.П., 1980, с. 36). Следовательно, всесторонний анализ сложного объекта обязательно включает в себя и качественные, и количественные моменты. При квантитативном анализе лингвистических объектов особое внимание следует уделять вопросам содержательного анализа как при квантификации материала (в частности при выделении единиц измерения), так и при интерпретации результатов анализа. Таким образом, всякая количественная характеристика лингвистических явлений предполагает качественную их характеристику. В то же время следует помнить, что качественная определенность лингвистического объекта, в свою очередь, существенным образом зависит от количества элементов, его образующих, от частоты употребления или от силы взаимодействия (корреляции) элементов. Можно констатировать наличие тесной взаимосвязи качественных и количественных характеристик языка; совместное их рассмотрение открывает широкие эвристические возможности исследования языковых процессов и явлений. Такой анализ допускает, например, прогнозирование качества со стороны количества и обратно. Открывается возможность изучения закономерностей корреляции квантитативных характеристик с качественными факторами (например, связь между частотностью и морфологическим строением слов). Во многих случаях количественные характеристики могут служить сигналом, направляющим внимание исследователя на некоторые скрытые от простого наблюдения качественные особенности и закономерности индивидуальных или функциональных стилей.

Можно сделать вывод, что квантитативное исследование языковых явлений, особенно в сочетании с системным подходом,

– не просто внешнее дополнение качественного анализа, а нечто большее, так как именно таким путем возможно более глубокое познание лингвистического объекта в его качественной определенности.

I.2. ЛИНГВИСТИЧЕСКИЕ ОСНОВЫ ИССЛЕДОВАНИЯ

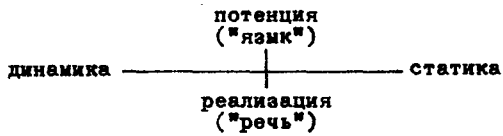
Ниже рассматриваются и конкретизируются некоторые теоретико-лингвистические аспекты квантитативно-системного исследования, связанные с разграничением языка и речи и расчленением системы речевой деятельности.

Язык и речь. Мы исходим из того общего положения, что разграничение языка и речи – это "разные способы интерпретации одного и того же материального объекта" (Леонтьев А.А., 1974, с. 44). Этот объект можно назвать "общей системой языка" или "речевой деятельностью", подчеркивая в последнем случае активный, деятельностный характер использования языка как главного орудия познания и общения в человеческом обществе. Речевая деятельность в широком смысле, которую для терминологического разграничения можно назвать "системой речевой коммуникации", включает кроме процесса порождения речи также процесс восприятия (принятия сообщений), который в данной работе не рассматривается.

Возможность и необходимость разграничения двух сторон речевой деятельности – "языка" и "речи" – покоятся на том очевидном факте, что в этой деятельности можно различить два взаимосвязанных, но все же ясно отделяемых друг от друга компонента: средство (орудие) и его применение. Однако уточнение содержания и значения этих компонентов осуществляется в лингвистической теории и практике по разным основаниям. Чаще всего выделяются парные признаки, такие как социальное – индивидуальное, идеальное – материальное, общее – отдельное, абстрактное – конкретное и др., которые, отдельно взятые или в комбинации, служат для определения и интерпретации понятий языка и речи. Характерно, что упомянутые пары признаков тесно соприкасаются друг с другом и частично накладываются друг на друга. В некоторых случаях они могут перекрещиваться. Отбор и комбинация этих признаков и вместе с тем разграничение сфер языка и речи проводится в зависимости от общего теоретико-методологического подхода и от особенностей конкретного научного исследования.

Учитывая специфику квантитативно-системного исследова-

ния языка вообще и лексики в частности, можно представить комплекс "язык – речь" в виде перекрещивания двух главных осей: оси с противопоставлением потенции – реализации и оси с противопоставлением динамики – статики:



Противопоставление потенции и реализации рассматривается в данном случае как собственно языковое-речевое соотношение, то есть, "языком" считается система потенциальных возможностей, а "речью" – актуализация этих возможностей в действительности. В соотношении потенции – реализации явно содержится момент иерархии в виде отношения следования, порядка, и поэтому представляется правомерным рассматривать это соотношение как уровневое, порядковое. Потенцию условимся считать высшим, "языковым" уровнем, а реализацию – низшим, "речевым" уровнем. Надо, однако, заметить, что такое разграничение в логическом плане не соответствует отношению первичности – вторичности в онтологическом плане, ср. общеизвестные высказывания о том, что "исторически факт речи всегда предшествует языку" (Соссюр Ф., 1977) и "именно в речи, реализованной в предложениях, формируется и оформляется язык" (Бенвенист Э., 1974, с. 140). Речь как конкретная реализация языка является единственным непосредственно наблюдаемым объектом лингвистики, и умозаключения о языковом уровне делаются на основе обобщения результатов анализа фактов речи. Однако в лингвистической практике возможен и "синтетический подход", когда осуществляется "выявление языковой системы через речь и исследование речи через языковую систему" (Ванников Ю.В., 1979, с. 17-18).

В квантитативной лингвистике и, в частности, при квантитативно-системном исследовании лексики противопоставление потенции и реализации имеет прямой практический смысл во многих отношениях.

Прежде всего с потенцией и реализацией (возможностью – действительностью) связана идея о "полной системе", т.е. полной группе событий, которые теоретически могут произойти в данных условиях, в противопоставлении к ограниченному набору действительно реализуемых событий. Квантитативно выраженное отношение между потенцией и реализацией может иметь эвристи-

ческое значение и выступать, например, как вполне осмысленный типологический критерий. Некоторые исследователи связывают потенцию и реализацию в указанном смысле с соотношением между статистической генеральной совокупностью и выборкой из этой совокупности и тем самым относят генеральную совокупность к языку, а выборку - к речи. Язык и речь можно в такой интерпретации соотнести с категориями общего и отдельного. Противопоставленность потенции и реализации используется и в теоретико-информационных исследованиях языка, в частности, при определении избыточности системы.

Далее, реализация может рассматриваться как актуализация, или выбор одного из возможных вариантов в данной ситуации. Например, выбор подходящего слова из данного семантического поля или актуализация одного из виртуальных значений слова в тексте. В этом смысле могут противопоставляться также словарь (лексикон) и текст.

Наконец, с понятиями потенции и реализации как характеристиками языка и речи связывают понятия вероятности и частотности. "Язык вероятностен, речь частотна" (Головин В.Н., 1968, с. 39). На этом основании различают, например, стили языка и стили речи: в первом случае учитываются стилевые вероятности (как языковая закономерность), а в другом случае - частоты появления единиц в отдельных текстах данного языка. Принимая вероятность за потенциальное, мы выделяем в ней "необходимое", а именно тенденцию частот (часто́ты как "случайное") группироваться вокруг некоторого постоянного, системно необходимого значения. Из этого следует, что "потенциальное, языковое" и "реализованное, речевое" могут быть при определенных условиях соотнесены с категориями необходимости и случайности.

Таким образом, мы видим, что выбранное нами противопоставление потенции и реализации, как коррелятов языка (орудия) и речи (применения), одновременно может быть соотнесено с такими философскими категориями, как возможность - действительность, общее - отдельное, необходимость - случайность. Все эти категории, частично накладываясь друг на друга, объективно характеризуют язык и речь и позволяют глубже осмыслить противопоставление и взаимосвязь двух основных аспектов речевой деятельности.

Что же касается соссюровского противопоставления социального (общественного) как признака языка и индивидуального как признака речи, то здесь возможен и другой подход, при

котором эти признаки рассматриваются как присущие одновременно и языку, и речи. Язык как потенция существует в "общей" форме, как инвариант для всех носителей данного языка; с точки зрения квантитативной лингвистики язык характеризуется в таком случае усредненными, обобщенными показателями в рамках данного языка. В то же время каждый индивид имеет свой индивидуальный язык – "внутреннюю систему" в противоположность к "внешней системе, или языку остальных" (Косериу Э., 1963, с. 300); в квантитативной лингвистике регистрируются соответственно вариантные особенности индивидуального языка, или идиолекта. Таким же образом, речь можно изучать в виде множества различных индивидуальных актов (сводных текстов) или на основе индивидуального творчества.

Наряду с осью потенции – реализации в нашей модели речевой деятельности выделяется и другая ось, которая перекрещивается с первой и тем самым разделяет уровни языка и речи (потенции и реализации) на подуровни, или сферы д и н а м и к и и с т а т и к и. Характеристики динамики и статики используются часто при разграничении речи как процесса и речи как результата этого процесса. Однако уровень языка рассматривается обычно в чисто статическом плане, как "инвентарь языковых средств и набор правил". Только в психолингвистике рассматривается специально и порождающий механизм, относимый к языковому уровню ("компетенция" в концепции Н. Хомского, "языковая способность" в трактовке А.А. Леонтьева). Разницу между сферами динамики и статики видят еще в том, что динамика (механизм и процесс) связывается с деятельностью мозга, в то время как статика (язык как "предмет") находится вне человека.

Исходя из нашего представления о перекрещивании осей потенции – реализации ("языка" и "речи") и динамики – статики и тем самым о совмещении их действия в речевой деятельности, представляется целесообразным различить динамический (порождающий, процессуальный) и статический (результативный, инвентарный) аспекты как на уровне языка, так и на уровне речи. В итоге, в системе речевой деятельности можно различить четыре основных аспекта (подсистемы): на языковом уровне – "языковую компетенцию" и "языковую схему", на речевом уровне – "речевой процесс" (акт) и "речевой продукт" (текст).

Иначе говоря, названные четыре аспекта, или подсистемы, выявляются вследствие того, что парные признаки потенция – реализация и динамика – статика накладываются друг на друга,

перекрещиваясь, в результате чего образуется структура, аналогичная четырехпольной таблице сопряженности альтернативных признаков. Соотношения между различными аспектами в пространстве перекрещивающихся признаков можно наглядно представить в виде таблицы (табл. I).

Таблица I

Аспекты речевой деятельности

Признаки	Динамика	Статика
Потенция ("язык")	Языковая компетенция	Языковая схема
Реализация ("речь")	Речевой процесс	Речевой продукт

Все названные аспекты речевой деятельности тесно взаимосвязаны, но для специфических исследовательских целей они могут быть изучены в отдельности. На основе различения сфер динамики и статики в квантитативной лингвистике можно выделить две основные о б л а с т и и с с л е д о в а н и я: работы, связанные с изучением процессов порождения речи, например, когда текст рассматривается как вероятностный процесс, и преобладающие до сих пор работы по вмяснению квантитативных характеристик речевого материала (текста) в статике. Соответственно, по линии восхождения от речевого к теоретическому, языковому уровню перед квантитативной лингвистикой стоят з а д а ч и двух типов: с одной стороны, объяснение фактов порождения речи, опираясь на языковую компетенцию, и, с другой стороны, извлечение из текстового материала "языковой схемы", т.е. статико-языковых закономерностей строения текста. Эти две задачи могут и должны быть объединены для полного представления и объяснения квантитативных сторон рече-языковой деятельности.

Ниже рассмотрим выделенные аспекты (подсистемы) речевой деятельности в отдельности. При этом мы начинаем с описания языковой схемы, как наиболее близкой к традиционному пониманию "системы языка" в статике (в рамках системного подхода следует признать системами и все другие части общей системы при их отдельном рассмотрении).

Языковая схема. Под названием языковая схема мы понимаем систему языковых элементов и отношений между ними. Эти отношения могут относиться как к парадигматике, так и синтагматике (Богданов В.В., 1973), они характеризуют грамматику

языка в широком смысле (как набор правил оперирования элементами). Языковую схему (ЯС) можно для наглядности представить в виде модели-формулы:

$$\text{ЯС} = \langle M; R \rangle, \qquad (I)$$

где M – множество элементов (инвентарь), R – совокупность системных отношений (грамматика). Языковая схема представляет собой таксономический феномен, в нее входят упорядоченные объекты, например, словарь и другие совокупности лексических единиц (лексико-семантические группы и т.п.) вместе с их формальными, семантическими, валентностными и др. отношениями и соответствующими квантитативными характеристиками.

Языковую схему можно рассматривать как статическую систему данного языка в целом, т.е. как суммарную совокупность лингвистических элементов и отношений между ними. Однако в реальности любой естественный язык существует во многих вариациях, в виде особых подсистем, относящихся к различным подъязыкам или функциональным стилям. Подъязык определяется как "набор языковых элементов и их отношений в текстах с однородной тематикой" (Андреев Н.Д., 1967, с. 23), то есть, понятие подъязыка связывается с определенной предметной сферой действительности (подъязыки публицистики, науки и техники, деловых документов и т.п.). Функциональные стили же определяются как "разновидности языка, обусловленные различиями в сферах общения и основных функций языка (общение, сообщение и воздействие)" (Виноградов В.В., 1967, с. 5–6). Исходя из данных определений, приходится констатировать, что понятия подъязыка и функционального стиля не совпадают. В определении функционального стиля подчеркивается стилистико-функциональный аспект языка (общение, сообщение, воздействие), в то время как подъязык рассматривается только с точки зрения сферы общения. Прямая соотнесенность между подъязыком и функциональным стилем может иметь место в том случае, если какой-нибудь подъязык характеризуется исключительно только одним, ему присущим стилем. Вопрос о соотношении подъязыков и функциональных стилей сложен и может быть решен только с помощью конкретных исследований на обширном материале. Во всяком случае, как тематические подъязыки, так и функциональные стили могут быть представлены для исследовательских задач в виде подсистем языка, обладающих своими специфическими (в том числе квантитативными) особенностями. Представляется даже возможным применить обобщенное название "подъ-

язык" для различных подсистем языка, выделенных на основе общего признака "сфера общения". С точки зрения квантитативно-системного исследования лексики важна констатация, что подъязыки отличаются друг от друга прежде всего "вероятностными спектрами своих лексических наборов" (Андреев Н.Д., 1967, с. 23).

Итак, языковая схема предстает как сложное, многоплановое явление, которое в зависимости от сферы общения в известной мере меняет свою структуру, т.е. выступает в различных вариантах. В то же время языковая схема отличается устойчивостью в рамках данного языка, в ней выделяется, например, общая лексика, присущая всем подъязыкам, и она характеризуется стабильными связями и закономерностями на общеязыковом уровне. Выработавшееся в общественной практике типичное и общепринятое языковое употребление, регулярно повторяющееся в определенной сфере коммуникации и отражающееся в языковой схеме в виде устойчивого "ядра" подсистемы, называется н о р м о й для данного языка или подъязыка (функционального стиля). В квантитативной лингвистике можно норму языка или подъязыка определить как наиболее вероятный состав элементов и наиболее вероятные отношения между элементами, а также такой доверительной диапазон значений измерения, который наиболее вероятен на уровне реализации в соответствующей совокупности текстов.+ Норма как статический феномен на уровне языковой схемы появляется в качестве регулятора в динамическом компоненте языковой компетенции (см. ниже).

Языковая компетенция. Языковая компетенция является следствием отражения сознанием людей языковой схемы, то есть, это – набор элементов и системных отношений между ними, сходных (в плане отражения) с элементами и отношениями языковой схемы, п л ю с особый динамический компонент, необходимый для приведения языка в действие. Модель языковой компетенции (ЯК) можно представить с помощью следующей формулы:

$$\text{ЯК} = \langle M'; R'; G \rangle, \qquad (2)$$

где M' и R' – отражение в сознании человека совокупности элементов и отношений языковой схемы, G – динамический компонент ("генератор"). Компонент G можно представить се-

+Наряду с нормой выделяется иногда узус, под которым обычно понимают неосознанную и некодифицированную норму. В нашей работе эти понятия не разграничиваются, и мы подразумеваем под термином "норма" оба эти понятия.

бе как механизм порождения в широком смысле, как сложный комплекс не только лингвистических умений (включая осознание "нормы"), но и знаний социально-прагматического порядка. Этот комплекс покоится отчасти на социально-коммуникативном опыте людей и отчасти на филогенетическом начале, связанном с особенностями устройства человеческого мозга. Здесь мы попадаем в "стыковую" область лингвистических, психолингвистических и психофизиологических исследований. Представляется, что без обращения к соседним наукам лингвистика не в состоянии объяснить действительную природу и внутренние закономерности речевой деятельности в целом. Точно так же квантитативные закономерности речи (универсальные особенности статистической структуры текста, наличие устойчивых распределений и регулярности в корреляциях с внешней средой и др.) могут быть объяснены только при условии привлечения к анализу экстралингвистических представлений и критериев.

Речевой процесс. Речевой процесс, или речевой акт, является реализацией потенций языка, то есть, самим действием механизма порождения речи. Отношение речевого процесса к языковой компетенции можно рассматривать также как отношение управляемой подсистемы к управляющей. Речевой процесс имеет своим непосредственным результатом линейную последовательность речевых единиц[+], но сам процесс порождения высказывания представляет собой сложный, многоступенчатый рече-мыслительный акт, состоящий из этапов превращения исходного замысла через внутреннюю речь в схему речевого высказывания и включения его в фонематические, лексико-семантические и логико-грамматические коды языка (см. Леонтьев А.А., 1969; Лурия А.Р., 1979).

Новым моментом, внесенным и з в н е в систему речевой деятельности, - объектом, над которым совершается действие, - оказывается исходный замысел, вызванный к жизни определенной потребностью и модифицированный мотивом и целью.[++] Назовем

[+]В зависимости от исследовательских задач в квантитативной лингвистике единицами речи можно считать звуки, слова, фразы и др. На лексическом уровне основной единицей следует считать слово, соответствующее минимальной смысловой единице в составе высказывания.

[++] Цель как предвосхищение результата действия формируется на основе мотива (включая личностную установку), за которым стоит потребность (внешний или внутренний стимул к деятельности). Различаются общие и промежуточные цели, например, промежуточной цели может соответствовать высказывание (выражение одной мысли), в то время как общей цели соответствует целое сообщение (выражение цельного комплекса мыслей), состоящее из отдельных высказываний.

этот модифицированный замысел т е м о й. Другим компонентом внесистемного (экстралингвистического) происхождения является с и т у а ц и я (обстановка, внешняя среда), т.е. условия, в которых происходит речевое общение. Ситуация охватывает сферу общения, а также сопутствующие случайные факторы (например, помехи), контекст в узком смысле и обратную связь. Тему и ситуацию можно считать действительными компонентами речевого процесса в той мере, в какой они появляются в отраженном виде в человеческом сознании (или подсознании) как элементы рече-мыслительного процесса. Здесь проявляются особенно ярко неразрывная связь речевого процесса с мышлением и взаимодействие с внешней средой, т.е. с окружающим миром.

В общей форме речевой процесс (РП) можно представить в виде модели-формулы:

$$\text{РП} = \langle T; S; G' \rangle, \qquad (3)$$

где T - тема (замысел, мотив, цель), S - ситуация (включая сферу общения, контекст, обратную связь и сопутствующие факторы), G' - языковой механизм в действии (смена состояний языковой компетенции).

Весь процесс порождения речи (порождения высказываний) представляет собой последовательность отдельных речевых актов, или действий. Речевое производство - это по существу циклический рече-мыслительный процесс, напоминающий в общих чертах функциональную систему поведения по П.К.Анохину (1962). Место этого процесса в общей системе речевой деятельности и его связанность с другими частями общей системы можно проиллюстрировать на схеме (см. рис. 1).

При рассмотрении отдельного акта порождения высказывания как рече-мыслительного действия выясняется, что этот акт (имеющий структуру РП в "микроформате") состоит из последовательности принятия решений, причем на отдельных стадиях приходится делать в ы б о р между различными возможными вариантами. Выбор варианта (например, слова на заключительном этапе рече-мыслительного действия) может в некоторых случаях быть вполне детерминированным данной ситуацией и возможностями (состояниями) языковой компетенции, но в других случаях говорящий действует в условиях "неполной определенности", когда известна лишь область ("поле"), где приходится искать подходящий вариант (например, при выборе слова из ряда синонимов). Говорящий опирается в таком случае на вероят-

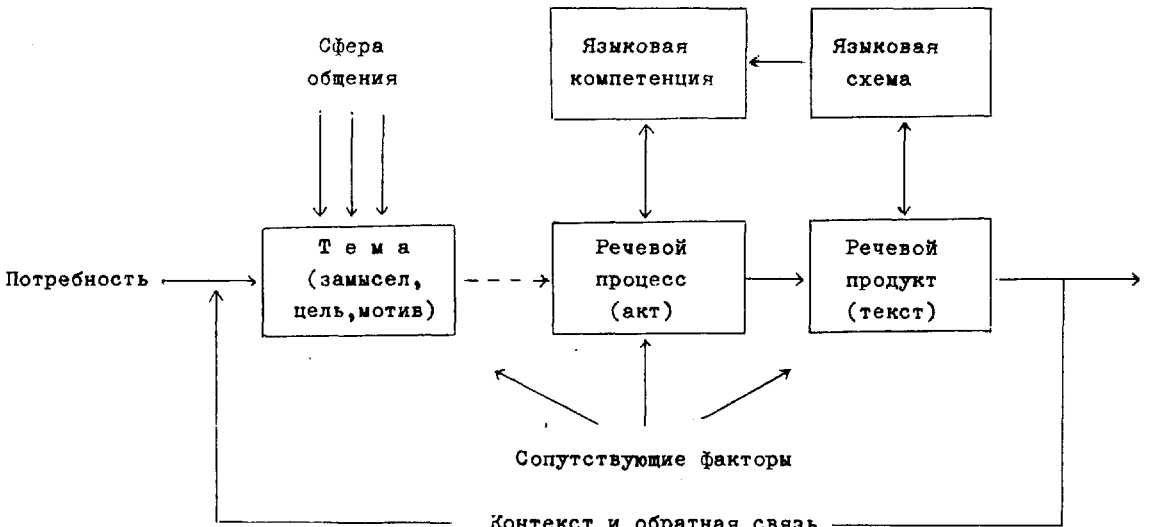


Рис. I. Модель системы речевой деятельности

ностное прогнозирование (Фрумкина Р.М., 1969; ср. также понятие вероятностной оценки у Налимова В.В., 1979), и в большинстве случаев выбирает самый типичный в данной ситуации вариант, причем возможны колебания в определенных границах. Вероятность выбора определенного варианта (например, слова) в этой системе обусловливается комплексом упомянутых факторов (тема, ситуация, состояние языковой компетенции). В итоге сам процесс порождения речи осуществляется в результате сложного взаимодействия детерминированных и вероятностных (случайных) факторов. Интегрально весь процесс можно рассматривать как вероятностную систему, характеризующуюся определенной устойчивостью и регулярностью в массе случайных событий.

При решении задачи конкретного математического моделирования процесса порождения речи пользуются обычно усложненными моделями марковского типа (например, Lounsbury F., 1965). В вероятностную схему порождения (и распознавания) текста можно ввести дополнительно ситуативные моменты (Пиотровский Р.Г., 1975, с. 56; Hoffmann L., Piotrowski R.G., 1979, с. 40–41). В некоторых моделях процесс порождения рассматривается явно как результат суперпозиции процессов случайных и детерминированных (Гачечиладзе Т.Г., Цилосани Т.П., 1971) или как результат наложения случайных связей на детерминированную информацию, причем отмечается, что диалектическое единство случайных и детерминированных связей позволяют системе развиваться и совершенствоваться (Зубов А.В., 1980).

Повышенный интерес к математическим, в частности, вероятностно-статистическим моделям порождения (и распознавания) речи обусловлен в наши дни практическими потребностями в такой области, как автоматическая переработка текста, а также в связи с решением некоторых задач лингвистического обеспечения систем искусственного интеллекта. В то же время внимание исследователей сосредоточивается на поисках теоретического объяснения квантитативных свойств порождения речи. В этой связи было высказано мнение, что адекватное математическое моделирование речевого процесса должно опираться на теорию оптимальных процессов, учитывая то, что (по Н.А. Бернштейну) "сама сущность речевой деятельности именно как деятельности заключается в оптимизации пути достижения поставленной цели по заданным параметрам" (Леонтьев А.А., 1974, с. 80).

Речевой продукт. Результатом процесса порождения речи является речевой продукт, или т е к с т, который в самом общем виде понимается как определенным образом (обычно письменно) зафиксированный отрезок речевого континуума. Текст на уровне реализации прямо соотносится с языковым уровнем как отображение языковой схемы в виде особой системы: $\langle M''; R'' \rangle$, где M'' и R'' - подмножества элементов и отношений между ними ($M'' \subset M$, $R'' \subset R$). По отношению к речевому процессу текст характеризуется тем, что в нем снята процессуальная форма рече-мыслительной деятельности, хотя он сохраняет связь с этой деятельностью. Будучи единственным прямо наблюдаемым объектом лингвистического исследования, текст может при соответствующем подходе послужить и моделью для изучения динамических сторон речи (т.е. речевого процесса). В этом смысле можно расширить понимание текста включением в его описание особым образом отраженных компонентов речевого процесса - темы и ситуации. Более того, при анализе текста можно пойти еще дальше и попытаться выявить в нем рефлексы языковой компетенции, воплощенные в тексте в виде определенных стилистических и прагматических черт. Следовательно, систему текста в целом можно представить следующей обобщенной моделью:

$$\text{Текст} = \langle M''; R''; T'; S'; G'' \rangle, \qquad (4)$$

где T' и S' - отражение темы и ситуации, G'' - отражение свойств языковой компетенции в тексте.

Учитывая то, что порождение речи (текста) рассматривается нами как сложный вероятностный процесс (формирующийся в результате взаимодействия детерминированных и случайных факторов), можно сделать вывод, что и результат такого процесса неизбежно должен характеризоваться вероятностными свойствами. В целом текст и соответствующий словарь могут быть рассмотрены как вероятностные системы. Эмпирически это проявляется, с одной стороны, в наличии стабильных распределений единиц, в устойчивых корреляциях внутрисистемных и межсистемных, в образовании "ядра" в замкнутых группах и т.п., а, с другой стороны, в явлениях периферии (результат размытости границ) и различного рода случайных флуктуациях.

Итак, на основе выделения перекрещивающихся осей потенции - реализации и динамики - статики общую систему речевой деятельности можно разделить на отдельные части, которые рассматриваются как основные подсистемы в плане функционирo-

вания системы. В то же время устанавливается внутренняя структура, взаимосвязь между отдельными подсистемами. Всю систему речевой деятельности можно для наглядности представить формализованно в виде "агрегатной" модели, состоящей из четырех частей:

$$\text{ЯС} = \langle M; R \rangle \qquad (1)$$

$$\text{ЯК} = \langle M'; R'; G \rangle \qquad (2)$$

$$\text{РП} = \langle T; S; G' \rangle \qquad (3)$$

$$\text{Текст} = \langle M''; R''; T'; S'; G'' \rangle \qquad (4)$$

Здесь выделяются следующие структурно-функциональные компоненты системы, выступающие в разных комбинациях в первичной или вторичной (отраженной) форме в составе подсистем: M - множество языковых элементов, R - набор отношений между элементами, G - порождающий механизм, T - тема, S - ситуация.

На заключительном этапе анализа системы речевой деятельности следует еще раз возвратиться к рассмотрению исходной модели языковой схемы. В соответствующем разделе говорилось о том, что языковая схема в реальности выступает в различных вариантах, которые можно называть подъязыками (или функциональными стилями). После того, как были проанализированы условия протекания речевого процесса и роль компонентов T (тема) и S (ситуация) в тексте, выяснилось, что эти компоненты (тема и ситуативный момент - сфера общения) связаны также с языковой схемой, и на основе сложного взаимодействия речевого и языкового уровней участвуют в расчленении языковой схемы на варианты на основании тематики и сферы общения. Исходную модель языковой схемы (см. формулу I) можно, таким образом, уточнить, дополняя ее факторами T'' и S'':

$$\text{ЯС} = \langle M; R; T''; S'' \rangle, \qquad (Ia)$$

где T'' и S'' рассматриваются как отражения типовых тем и сфер общения при образовании вариантов языковой схемы (подъязыков или функциональных стилей). На том же основании можно, например, различить в языковой схеме стилистические пласты лексики.

Безусловно, представленная модель речевой деятельности отражает действительность лишь в определенном аспекте, упрощенно. Дальнейшее развитие модели связано как с детализацией ее отдельных сторон (в зависимости от задач исследования), так и с углубленным изучением внутри- и внесистемных связей.

Моделирование общей системы языка и изучение свойств и

взаимосвязей отдельных частей и компонентов системы позволяет четко разграничить предмет и сформулировать задачи лингвистического исследования. В то же время вся система предстает перед исследователем как целостное образование. В соответствии с этими установками ведется и конкретное исследование. Например, квантитативно-системное исследование лексики может вестись в плане динамики и статики или в их взаимосвязи. Можно исследовать текст как последовательность грамматически и семантически связанных лексических единиц (на уровне $\langle M''; R'' \rangle$) или текст как связное целое с темой и стилистическими или прагматическими особенностями (добавляя компоненты $T'; S'; G''$). Выявление частотных свойств текстов и соответствующих словарей, установление системных связей между словами в парадигматике и синтагматике, изучение фактов порождения текста и т.д. должны вести к последующему обобщению, упорядочению и осмыслению эмпирического материала на более высоком, теоретическом уровне. В конечном счете следует осуществить синтетический, интегральный подход к изучению и объяснению квантитативных свойств системы речевой деятельности в целом и в неразрывном единстве квантитативного анализа с качественной интерпретацией.

I.3. МЕТОДИКА ИССЛЕДОВАНИЯ

В данном разделе освещаются вопросы, относящиеся к методике исследования: квантификация материала, единицы и уровни анализа, приемы описания лексического материала в виде лексических групп и моделирование с помощью распределений; особое внимание уделяется интерпретации лингвистических распределений.

Статус методики. Под методикой исследования мы понимаем упорядоченную совокупность, или с и с т е м у, отдельных методов (конкретных способов или приемов решения задачи), а также применение этой системы. Конкретные методы исследования находятся с точки зрения методологии научного знания на уровне "методики и техники исследования" (Садовский В.Н., 1979); этот уровень является низшим в иерархии, включающей более высокие уровни - конкретно-научной, общенаучной и философской методологии. Отнесение методики и техники исследования к "низшему" уровню не означает, однако, что на этом уровне разрывается связь с высшими уровнями, в данном случае

с общими теоретико-методологическими принципами квантитативно-системного анализа (уровни конкретно-научной и общенаучной методологии) или с уровнем философской методологии. Применяемые методы должны в своей совокупности образовать такую систему научно разработанных правил и исследовательских приемов, которая входила бы как органическая часть в более общую систему научной методологии.

Квантитативно-системный анализ (как и всякий системный анализ) характеризуется не специфическим аппаратом конкретных методов, а упорядоченным, логически обоснованным подходом к использованию существующих методов, которые уже разработаны в рамках других наук (математика, лингвистика и т.д.). Выбор и комбинация этих методов представляют собой один из важнейших моментов как на этапе наблюдения и эксперимента, так и на этапе анализа и теоретического обобщения результатов исследования. Следует подчеркнуть важность разработки методики в исследованиях "на стыке наук", в том числе в работах по квантитативной лингвистике, где выбор адекватных методов может стать важным шагом на пути к обнаружению новых явлений и зависимостей в лингвистической науке.

В рамках представления и описания применяемых в данной работе методов квантитативно-системного анализа лексики необходимо прежде всего осветить некоторые общие положения, связанные с квантификацией лингвистического материала как эмпирического базиса исследования. При освещении проблем методики исследования мы не будем рассматривать отдельные конкретные методы квантитативно-лингвистического анализа, которые достаточно подробно описаны в соответствующих учебниках и специальных исследованиях[+], а остановимся прежде всего на логических основах и общих принципах применения квантитативных методов в лингвистике.

Квантификация, квантование, измерение. Предпосылкой применения квантитативных методов при изучении языковых явлений является к в а н т и ф и к а ц и я исследуемого материала. Квантификация – в широком смысле слова[++] – означает

[+] См., например, Головин Б.Н., 1971; Бектаев К.Б., Пиотровский Р.Г., 1973-1975; Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А., 1977; Piotrowski R.G., Bektaev K.B., Piotrowskaja A.A., 1985; Altmann G., 1980a; Dugast D., 1980.

[++] В более узком смысле (в логике) под квантификацией понимается точное выявление и определение объемов субъекта и предиката суждения, а также применение к логическим выражениям особых операторов, именуемых кванторами (Кондаков Н.И., 1971, с. 211).

количественное представление качественных явлений, т.е. такую процедуру, при которой исследуемым явлениям, имеющим принципиально качественную природу (например, лингвистическим объектам), приписываются количественные оценки, вследствие чего эти явления могут быть изучены как количественные объекты. Как правило, квантификация предполагает (включает) предварительное "квантование" объекта (приведение объекта в удобную для измерения форму, выделение единиц учета на разных уровнях рассмотрения) и последующее измерение. Таким образом, в рамках методики квантитативно-лингвистического исследования мы можем определить квантификацию в целом как некоторое абстрактное преобразование (сведение качества к количеству), реализуемое посредством квантования и измерения.

Как уже было отмечено, к в а н т о в а н и е представляет собой предварительный этап квантификации, необходимый в тех случаях, когда единицы учета явно не даны в наблюдении или когда по условиям исследования приходится их каким-то образом модифицировать. Например, так называемые непрерывные переменные величины могут быть приведены к дискретному виду путем их квантования, т.е. путем разбиения области их изменения на совокупность непересекающихся интервалов. Естественно, что всякие процедуры квантования лингвистических объектов должны быть основаны прежде всего на содержательных соображениях и быть однозначным образом интерпретируемы.

И з м е р е н и е можно определить как процедуру, приписывающую по заранее фиксированным правилам числовые значения тем или иным наблюдаемым объектам. Можно показать, что измерение имеет своим логическим основанием категории свойства и отношения. При измерении выявляется отношение объекта (X) на основе какого-либо свойства (P) к значению (У), выраженному количественно. Эту абстрактную структуру можно представить в виде формулы-схемы:

$$X \xrightarrow{P} У, \qquad (I)$$

где компонент "$\xrightarrow{P}$" выражает "отношение отображения"[+], при котором объект (X) соотносится со значением (У) на основе квантитативного свойства (P), т.е. такого свойства, которое

[+] Ср. высказывание А.И. Ракитова (1977, с. 240) о сущности измерений: "Измерения по существу выступают как определенные функции, осуществляющие изоморфные или гомоморфные о т о б р а ж е н и я (разрядка наша. - Ю.Т.) элементов, ситуаций, процессов или отношений одной системы /.../ в элементы другой - числовой - системы".

допускает количественную оценку. Квантитативными свойствами являются, например, объем, количество, частота, протяженность и др. Примером, иллюстрирующим схему (I), может служить следующее утверждение: текст (X) связан с количественно выраженным значением (У) в том смысле, что объем текста (Р) равен этому значению. В более привычной форме это утверждение можно выразить таким образом:

объем (Р) текста (X) равен У.

Другие примеры такого же рода:

частота (Р) слова (X) равна У;

длина (Р) слова (X) равна У.[+]

Эти примеры показывают, что, говоря о количественном представлении объекта на основе какого-либо свойства, мы в сущности приписываем количественное значение не самому объекту, а свойству, или п р и з н а к у объекта. Притом понятие признака (параметра, характеристики) можно считать "материализацией" логического понятия свойства на уровне наблюдения и эксперимента. В практической работе удобно различать "наименование признака" (Р) и "значение признака" (У), как наиболее существенные компоненты рассматриваемой логической схемы (I), и результат измерения, т.е. значение признака (У) можно рассматривать как функцию:

$$У = f\left[P\,(X)\right] \quad , \qquad (2)$$

где Р(X) - признак (наименование признака) объекта X.

В отношении понятия "значение признака" (У) следует уточнить, что оно состоит, как правило, из двух элементов: из количества (числа) и наименования единиц измерения, принятых для данного признака, например: "длина слова равна 6 буквам". Можно также указать на то, что наряду с чисто количественными (числовыми) значениями признаков существует и смешанные, или "промежуточные" значения признаков (Рубашкин В.Ш., 1976): оценочно-количественные значения (например: очень мало, мало, средне, много, очень много) и оценочно-бинарные (имеется - не имеется; много - немного и т.п.). Оценочные значения могут использоваться в определенных условиях

[+] Примечательно, что идентичную логическую структуру (1) имеют высказывания типа "яблоко - красное", т.е. "яблоко" (X) связано с "красное" в том смысле, что "цвет" (Р) яблока равен "красное" (У) (ср. Кохонен Т., 1980, с. 15). В изоморфизме структур представления качественных и количественных значений свойств объектов просматривается глубокая взаимосвязь категорий качества и количества.

(при "нечетких" множествах) или на определенных этапах анализа; при необходимости они могут быть заменены числовыми знаками, например, когда оценочно-количественные значения выражаются через "баллы" (образующие порядковую шкалу) или когда бинарным отношениям приписываются I и O (образующие т. наз. дихотомическую шкалу).

Далее можно указать на то, что кроме п е р в и ч н ы х (прямых) измерений выделяются еще п р о и з в о д н ы е (косвенные) измерения. Измерения называются первичными, если они не основываются на каких-то предварительных измерениях; в противном случае они называются производными (Суппес П., Зинес Дж., 1967, с. 25). В качестве классического примера производных измерений можно привести измерение "скорости" движения, определяемое как отношение длины пути и промежутков времени. В квантитативной лингвистике производными измерениями считаются, например, "индекс разнообразия" (отношение объема словаря к объему текста); "коррелятивная функция" как отношение условной вероятности к независимой вероятности (Андреев Н.Д., 1967, с. 22) и многие другие функции или отношения частот (например, стилеметрические коэффициенты Б.Н. Головина, 1971, с. 140-154).

Аналогично различению первичных и вторичных измерений можно разделить и квантитативные свойства на первичные, или п р о с т ы е, и вторичные, или с л о ж н ы е. Последние характеризуются комплексной структурой. Как показывает опыт, в практической работе очень часто приходится иметь дело именно со сложными свойствами (признаками). В отношении лингвистических объектов можно сказать, что они могут характеризоваться множеством различных (первичных и вторичных) квантитативных свойств, причем задачей квантитативно-системного исследования лексики можно считать выявление и осмысление "системных свойств", т.е. таких свойств, которые могут служить основанием для обнаружения в лексике специфических квантитативно-системных закономерностей.

С точки зрения теории и практики измерений важно еще учитывать то, что измерения могут производиться по разным уровням, или ш к а л а м (номинальная, порядковая, интервальная и пропорциональная шкалы). Каждая шкала характеризуется соответствующей числовой системой и возможными (допустимыми) эмпирическими операциями (Стивенс С.С., 1960).

Единицы и уровни анализа. При квантитативном исследовании любой системы обязательно требуется выявить единицы анализа, которые поддаются подсчету как повторяющиеся компоненты (элементы) данной системы на данном уровне исследования.[+] Условие повторяемости единиц естественным образом связано с требованием инвариантности этих единиц. При изучении лингвистических объектов это означает, что одинаковые языковые единицы в разных их проявлениях должны отождествляться друг с другом, чтобы можно было констатировать факт их повторяемости.

Если обычно с л о в о рассматривается как основная лексическая единица (ЛЕ), т.е. единица лексической системы языка, то при квантитативном подходе требуется конкретизировать это понятие как единицы исследования. Следует различать два аспекта существования и функционирования слова в речевой деятельности - как единицы с л о в а р я и как единицы т е к с т а. Этим подчеркивается то обстоятельство, что исследование лексики охватывает изучение словарного состава не только как структурированной совокупности лексических единиц, но и как некоторой функционирующей коммуникативной системы. Остается уточнить понятия словаря и текста с точки зрения их квантитативных свойств. На лексическом уровне эти понятия определяются следующим образом.

Под словарем (словником) понимается совокупность различных слов, представляемых обычно в виде списка. Единицы словаря могут быть двоякого рода. В одном случае мы можем составить список с л о в о ф о р м (СФ), т.е. слов в таком виде, как они встречаются в реальных текстах. В другом случае мы объединяем разные формы под общим знаменателем, обычно под так называемой основной формой (у существительных - форма именительного падежа, у глаголов - инфинитив). Такие единицы словаря называем л е к с е м а м и (ЛК).

Под текстом мы понимаем в общем случае линейную последовательность каких-то определенных языковых единиц: слов,

---

[+] Понятия "единица" и "элемент" можно считать близкими, с той разницей, что понятие "единица" обязательно включает в себя момент инвариантности, необходимый для установления определенных квантитативных свойств, подлежащих измерению, а понятию "элемент" (в терминологии системного исследования) присущ момент взаимосвязи с другими элементами, которые в своей совокупности образуют целостную систему. Общей для единиц и элементов является неразложимость (элементарность) на данном уровне расчленения системы.

морфем и т.д. На лексическом уровне текст представляется как линейная совокупность единиц, именуемых в обыденной речи словами, а в квантитативной лингвистике – *словоупотреблениями* (СУ). Формально их определяют как промежутки между двумя пробелами в тексте.

Упомянутые термины – словоформа, лексема, словоупотребление – строго различаются на определенных этапах квантитативного анализа лексики. В тех случаях, когда различение этих терминов несущественно, употребляется общее название *слово*.

Приведенное выше определение единиц словаря и текста относится к тому случаю, когда слово идентифицируется только по своей внешней форме или как единство формы и лексического значения. С точки зрения требования инвариантности единиц (на основе которой можно идентифицировать эти единицы и установить факт их повторяемости) можно, однако, представить себе и другой случай, когда критерием различения считается отдельное значение слова и единицей подсчета берется соответствующий лексико-семантический вариант (ЛСВ). Таким образом, в зависимости от задач исследования единица анализа – слово – может рассматриваться как с точки зрения плана выражения, так и плана содержания.

В некоторых случаях единицами лексики могут быть слова в качестве представителей определенных классов, например, "односложные слова, двусложные слова...", "существительные, глаголы ...", которые идентифицируются и подлежат подсчету по признаку определенных фонетических, морфологических и др. характеристик. Другими словами, критерием идентификации слова как единицы анализа является принадлежность его к *классу слов*, заданному некоторым свойством (качественным или количественным). Такими свойствами могут быть, например, образование слова по данной модели, принадлежность к определенной части речи, лексико-семантической группе и т.д. Свойства, на основе которых слово может быть идентифицировано как единица анализа, определенным образом связаны с различными уровнями языка. Отсюда можно заключить, что уровень анализа (описания) явлений лексики, т.е. лексический уровень в целом подразделяется на *подуровни*, например, фонетико-лексический и грамматико-лексический (под)уровни наравне с собственно лексическим, или "лого-лексическим" (словесным) уровнем. Во всех этих случаях можно различать планы выражения и содержания, хотя

наиболее ярко различие между этими планами выступает на собственно лексическом уровне (различаются лексико-формальный и лексико-семантический аспекты). В задачи квантитативно-системного исследования лексики может входить также исследование стилистических свойств лексики, которое образует особый лексико-стилистический аспект анализа.

Итак, с точки зрения различных аспектов анализа слово выступает как единица, которая приобретает каждый раз особое качество, определяемое интересующими исследователя свойствами. Можно констатировать, что исследование лексики в разных аспектах (на разных подуровнях) представляет собой по существу ряд исследований различных п о д с и с т е м лексики, причем единицей (и элементом) этих подсистем неизменно выступает слово в разных своих проявлениях.

В то же время само слово может рассматриваться как особая "система форм и значений" (Виноградов В.В., 1947, с.15), то есть как отдельный системный объект, состоящий из элементов и имеющий определенную структуру. Слово рассматривается в таком случае с точки зрения его внутреннего строения, причем единицей учета оказывается уже не само слово, а составляющие его элементы: фонемы, морфемы, слоги и т.д. Квантитативное исследование внутреннего строения слова относят обычно к областям фоно- или морфостатистики, но такое исследование все же тесно связано с лексическим уровнем анализа в том смысле, что оно представляет собой необходимый предварительный этап на пути к исследованию лексики на фоно- и морфолексических подуровнях, когда единицей подсчета оказывается слово как представитель класса. Кроме того, одной из задач квантитативно-системного анализа лексики является установление взаимодействия разных уровней лингвистического анализа, например, выявление фонологических и лексических соответствий (корреляций) или сопоставление словообразовательных и смысловых структур слова в их квантитативных соотношениях. Исследование межуровневых связей составляет один из важных методологических принципов системного анализа ("связь системы со средой"), оно имеет также значение для решения некоторых общих теоретико-лингвистических задач, например, в области типологического изучения языков.

Лексические группы. Системный характер лексики обнаруживается особенно ярко в объединении слов в различные группы (классы, ряды, поля и т.п.) на основании обнаруженных сходств и различий между словами. Образование лексических группиро-

вок основывается на процедуре классификации и во многом зависит от метода и целей исследования. В этом смысле возможны разные решения задач классификации на одном и том же материале. Создаваемые исследователем группировки следует рассматривать как построения, позволяющие моделировать структуру словаря с какой-нибудь определенной точки зрения. Но это не значит, что различные лексические группировки лишены онтологического основания, при условии, что учитываются реально существующие связи в лексической системе языка.

Группировка слов и изучение внутригрупповых и межгрупповых отношений - это чрезвычайно полезный прием лексического анализа. Во многих случаях в результате такого анализа удается выявить еще не описанные стороны лексической системы языка и организовать их дальнейшие исследования. Общеизвестна роль лексических группировок (классификаций) в лексикографии, стилистике и методике преподавания языков, а также при информационном поиске и автоматическом переводе.

Сначала требуется уточнить некоторые понятия и термины. В зависимости от характера взаимосвязи слов мы выделяем в системе лексики особые подсистемы под общим названием "лексические группы" (ЛГ), которые имеют свои разновидности на разных уровнях описания лексики. На основании сходства фонетического состава или словообразования выделяются "лексико-формационные группы" слов (ЛФГ). Общность грамматических значений является основанием для образования "лексико-грамматических групп" (ЛГГ). На основании близости или однородности семантики слов образуются "лексико-семантические группы" (ЛСГ). Кроме этих основных типов ЛГ, возможны еще группировки слов по разным другим основаниям (этимологическим, стилистическим и др.). Возможна группировка и по смешанным признакам.

Лексические группы, образуемые на любом уровне и с помощью любой классификационной процедуры, обладают весьма существенными к в а н т и т а т и в н ы м и с в о й с т в а м и. Например, в пределах обследуемого конечного словаря можно определить объем, т.е. количество элементов ЛГ, причем на конкретном уровне численные значения объемов различных ЛГ образуют типичное многообъектное распределение (см. ниже). При неодинаковой величине объемов ЛГ такое распределение может быть представлено как ранговое, выявляя упорядоченность по "значимости" в ряду лексических групп и тем самым раскрывая некоторую лингвистическую закономер-

лость. Далее, при употреблении в речи выявляются частоты отдельных ЛГ, а также внутригрупповые распределения частот элементов в рамках ЛГ, что позволяет определить центр (ядро) и периферию в данной совокупности объектов. Учитывая "размытость" границ лексических групп (возможность пересечения), а также вероятностный характер распределения частот употребления слов (возможность выбора), можно считать правомерным при исследовании ЛГ применить наряду с качественным анализом квантитативно-системный подход, при котором распределения квантитативных характеристик ЛГ как в словаре (инвентаре), так и в тексте (речи) рассматриваются в качестве вероятностных систем с характерными для таких систем свойствами устойчивости и вариативности. Притом основным инструментом квантитативно-системного анализа лексики, в том числе ЛГ, служит моделирование с помощью распределений, которому посвящается следующий раздел.

Моделирование с помощью распределений. Моделирование, т.е. построение и анализ моделей (аналогов реального объекта - оригинала) является методом системного изучения внутреннего строения объекта или его поведения. Построение модели - это, по существу, попытка проникнуть в природу и "архитектонику" связей исследуемой системы. Известно, что модели могут иметь различную природу (см., например, Штофф В.А., 1972). Для нас важно установить принципы построения и анализа таких моделей, которые существенным образом связаны с квантитативным исследованием лексики как вероятностной системы. Этому отвечает в первую очередь моделирование с помощью р а с п р е д е л е н и й, учитывая то, что распределение как обобщающее, интегральное понятие является важнейшей структурной характеристикой именно вероятностных систем. "При помощи понятия распределения, - отмечает Ю.В. Сачков (1971, с. 112), - характеризуются элементы вероятностных систем, их взаимосвязь и основания их вхождения в системы и сами системы в целом." Важно не только то, что распределение выражает наличие внутренней упорядоченности в системе, но и то, что оно отражает взаимодействие между элементами и общность в их поведении, т.е. целостность системы, а также устойчивость и регулярность в массе вероятностно-случайных явлений. Учитывая то, что распределение в значительной мере определяется внутренними свойствами элементов системы, можно на его основе изучать и свойства отдельных элементов (например, слов или

классов слов). Распределение представляет собой, таким образом, настоящий синтез внутренней расчлененности и общности в строении системы.

Само понятие распределения можно понимать в широком смысле, как упорядоченную совокупность количественно выраженных значений, т.е. результатов измерения объекта (объектов), обычно с указанием значимости (частоты, вероятности, ранга) этих значений в данной совокупности. В более узком смысле (вероятностное) распределение определяется как "перечисление значений случайной величины и их вероятностей" (Венецкий И.Г., Кильдишев Г.С., 1975, с. 110). Но и в таком случае оно предстает как широкое понятие, например, под определение вероятностного распределения попадает и единичная вероятность (т.к. установление вероятности А возможно лишь при одновременном определении вероятности не-А в данной совокупности). Таким образом, распределение как в широком, так и в узком смысле охватывает широкий круг явлений при квантитативном исследовании объектов; оно может включать простейшие формы измерений (один элемент в данной совокупности) и более сложные формы соотношений между измерениями. Главное здесь то, что само распределение должно интерпретироваться системно, в данном случае с точки зрения структурных характеристик вероятностных систем (взаимосвязь элементов при наличии моментов устойчивости и вариативности).

Для выявления основных возможностей представления (моделирования) данных квантитативного исследования лексики в виде распределений целесообразно взять за основу методологический принцип разграничения уровней я з ы к а (потенции) и р е ч и (реализации) и сфер с т а т и к и и д и н а м и к и в рамках общей системы речевой деятельности (см. гл. 1.2).

Учитывая выделение в системе речевой деятельности уровней языка и речи, распределения могут быть разделены на т е о р е т и ч е с к и е (языковые) и э м п и р и ч е с к и е (речевые).

Как уже было сказано, наряду с разграничением уровней языка и речи (уровней потенции и реализации) в системе речевой деятельности различаются сферы статики и динамики. Соответственно распределения можно разделить на с т а т и ч е с к и е и д и н а м и ч е с к и е. Это различение имеет в первую очередь содержательный смысл, так как с точки зрения внешней формы они не отличаются друг от друга.

Статические распределения, называемые также "синхронными", выражают преимущественно синхронический и парадигматический аспекты анализа лингвистических явлений. Сюда относятся, например, ранговое распределение слов или распределение частот слов в тексте ("частотный спектр" лексики). В то же время в качестве статических распределений можно рассматривать разбиения некоторых совокупностей (например, лексико-семантических полей) на отдельные группы (классы, кластеры), упорядоченные с точки зрения их вероятностно-частотных свойств.

Динамические распределения в лингвистике отличаются от статических распределений тем, что они выражают либо процессуальность, связанную с явлениями порождения речи (в синхронии), либо изменение, развитие языка (в диахронии). Во всех случаях явно или неявно учитывается момент времени, как изменяющийся фактор в составе распределения, и поэтому динамические распределения часто именуются также "диахронными" в широком смысле этого слова.

Таким образом, кроме разделения распределений на теоретические (языковые) и эмпирические (речевые) выделяются характерные для лингвистики статические (синхронные) и динамические (диахронные) распределения. Наряду с этими основными типами лингвистических распределений можно различать еще ряд подтипов, или разновидностей распределений, которые будут рассмотрены ниже.

Разновидности распределений. С точки зрения технического оформления всякое распределение может быть задано в виде таблицы ("ряда распределения"), графика или может быть выражено математически в виде формулы (функциональной зависимости). Притом распределение может быть представлено как д и ф ф е р е н ц и а л ь н о е или и н т е г р а л ь н о е (кумулятивное) в виде соответствующих таблиц, графиков или функций (различаются дифференциальная функция, или "функция плотности", и интегральная функция; см. Митропольский А.К., 1971, с. 209).

В силу некоторых теоретических и практических соображений д и с к р е т н о е распределение лингвистических данных обычно описывается через н е п р е р ы в н у ю функцию распределения (Бектаев К.Б., Пиотровский Р.Г., 1973, с.131).

Не входя в детали применения конкретных технических приемов предварительного упорядочения и статистической оценки материала (сводка, группировка, вычисление характеристик рас-

сеяния и т.д.), мы остановимся на тех вариантах представления данных в виде распределений, которые соответствуют рассмотренным выше основным правилам квантификации лингвистического материала. Мы исходим из формул измерений (I) и (2), где соотносятся друг с другом три основных компонента: объект (X), признак (P) и значение признака (У). Результаты измерений можно на этом основании представить в виде таблиц, которые являются основой для выявления соответствующих распределений. Мы различаем три основных схемы при построении таблиц - рядов распределения.

Схема I: о д н о о б ъ е к т н о е распределение[+] (один объект, несколько признаков)

	P_I	P_2	...	P_n
X	$У_I$	$У_2$	...	$У_n$

По этой исходной схеме один объект соотносится с результатами измерения в разных условиях. Например, если X - конкретная лингвистическая единица (или класс этих единиц), P_i (наименование признака) - частота в тексте i; $У_i$ - конкретные значения частот; то удается проследить "поведение" одной определенной лингвистической единицы в серии испытаний. Если ранжировать данные по убыванию или возрастанию количественных значений $У_i$ (и соответственно P_i), то можно:

а) составить вариационный ряд с указанием частот или вероятностей появления $У_i$; в этом случае образуется т.н. спектр, или спектральное распределение;

б) приписать ранги (порядковые номера) значениям $У_i$, в результате чего получается т.н. ранговое распределение.

По спектральному однообъектному распределению определяется вид распределения (в отношении лингвистических объектов это чаще всего распределение "гауссового семейства", т.е. биномиальное, нормальное, пуассоновское и др.). Ранговое распределение при данной схеме однообъектных распределений обычно не представляет интереса при изучении лингвистических объектов. Особый случай возникает при ранжировании (упорядочении) $У_i$ по качественным соображениям, на-

[+] Однообъектное распределение соответствует "горизонтальному" распределению по терминологии П.М. Алексеева и "однопредметному" распределению по Г.Я. Мартыненко (1982), хотя имеется принципиальная разница в подходе к решению проблем типологии лингвистических распределений.

пример, при рассмотрении динамических (диахронных) процессов. В таких случаях можно исследовать распределение как "тренд" - в виде функциональной зависимости значений y_i от $P(t_i)$, где t_i - моменты времени.

Схема 2: м н о г о о б ъ е к т н о е распределение[+] (один признак - несколько объектов)

	P
X_1	y_1
X_2	y_2
⋮	⋮
X_m	y_m

Согласно этой схеме разные объекты измеряются по одному общему признаку. Например, X_i - разные лингвистические единицы (или классы единиц), P - частота в одном тексте, y_i - конкретные значения частот. По этой схеме составлен, например, обыкновенный частотный словарь слов или классов слов (частей речи, фонетических, морфологических, семантических и др. типов слов), а также любой частотный список разных других лингвистических единиц. При этом получается распределение частот разных единиц относительно друг друга в данной совокупности (например, в данном тексте).

Здесь, так же как и при однообъектном распределении, различаются две разновидности (формы представления):

а) спектральное распределение (спектральная форма распределения), когда одинаковые результаты измерений объединяются в группы с указанием числа объектов с данным результатом измерения; например, когда исследуется зависимость между частотой слова в тексте и количеством слов с данной частотой (т.н. частотный, или лексический спектр);

б) ранговое распределение, при котором ранжированным значениям частот y_i приписываются ранги (i) и исследуется зависимость между y_i и i (например, ранговое распределение частот слов).

Как спектральное, так и ранговое многообъектное распределения в лингвистике относятся, как правило, к распре-

---

[+] Ср. "вертикальное распределение" по П.М. Алексееву (1978 и 1985) и "многопредметное распределение" по Г.Я. Мартыненко (1982).

делениям "негауссового семейства". В них проявляется одно из характерных свойств коммуникативных систем: асимметричное распределение элементов по "значимости", вследствие чего основную функциональную нагрузку несет небольшое число доминирующих элементов ("ядро").

Схема 3: к о м п л е к с н о е (многомерное) распределение (несколько объектов - несколько признаков)

	P_I	P_2	...	P_n
X_I	y_{II}	y_{I2}	...	y_{In}
X_2	y_{2I}	y_{22}	...	y_{2n}
...				
X_m	y_{mI}	y_{m2}	...	y_{mn}

Данная схема комбинируется из двух первых схем. В простейшей форме число объектов (X) или число признаков (P) равняется двум, например, когда сравниваются частоты разных слов в двух разных текстах или частоты разных классов слов в двух аспектах: в словаре и тексте. На основе схемы комплексного распределения исследуются "многомерные" задачи: взаимосвязь и совместная вариация ряда объектов или признаков. Связь между распределениями количественных значений (по горизонтали или вертикали) может быть выражена функциональной зависимостью (уравнением регрессии), а сила этой связи может быть измерена коэффициентом корреляции (линейной или нелинейной). Внутренние связи во всей совокупности данных можно установить методами факторного анализа, кластер-анализа и др.

Отличительной чертой данного подхода к классификации (типологии) лингвистических распределений является то, что здесь исходят из качественных представлений о распределении, как неком результате измерения, при котором соотносятся друг с другом три основных компонента: объект (X), признак (P) и значение признака (У). Хотя в строгом смысле собственно распределением можно считать лишь ряд У в спектральной или ранговой форме, все же при конкретном анализе нельзя забывать о "происхождении" этого ряда, т.е. о связи компонента У с компонентами X и P. Это необходимо не только для выработки предварительных гипотез (о форме распределения и т.д.), но и для содержательной интерпретации резуль-

татов квантитативного анализа с учетом специфики и задач л и н г в и с т и ч е с к о г о исследования. Кроме того, как было показано выше, именно учет всех компонентов процедуры измерения (X, P и У) позволяет наиболее четко и естественно различать основные типы лингвистических распределений: однообъектных ("горизонтальных"), многообъектных ("вертикальных") и комплексных.

На практике допускается "гибкая" трактовка понятий объекта, признака и значения признака. Объектом может быть индивидуальная единица (например, конкретное слово) или класс единиц (часть речи и т.п.). Признак (точнее, наименование признака) может быть простым или сложным, например, "частота в тексте T_i". Значение признака может быть количественно-пропорциональным, интервальным или порядковым (в т.ч. "балльным"). Вследствие такой гибкости в определении компонентов измерения возможны различные варианты классификаций лингвистических распределений в зависимости от конкретных условий и задач исследования.

Итак, все рассмотренные разновидности распределений могут служить квантитативному исследованию и моделированию лингвистических объектов. При этом возможны различные способы представления и описания распределений, в том числе в виде функциональных зависимостей. Вполне возможно также описание некоторых типов распределений в терминах теории нечетких (размытых) множеств (Заде Л., 1976; Лесохин М.М. и др., 1982; о возможности вероятностного подхода к теории нечетких множеств см. Налимов В.В., 1979а). При исследовании некоторых динамических явлений (например, процессов порождения речи) могут быть полезны и специфические методы исследования вероятностных (случайных) процессов с привлечением теории оптимальных процессов, теории "динамического программирования" (Вентцель Е.С., 1976) и др.

Ориентация на моделирование с помощью распределений при квантитативно-системном подходе объясняется спецификой системного исследования, которая состоит в том, что изучение объекта осуществляется именно в том аспекте, в каком он представляет систему при данном подходе. В этом смысле лингвистическое распределение является м о д е л ь ю - о п и с а н и е м тех языковых объектов, которые можно представить себе как вероятностные системы.

Исследование лингвистических распределений начинается, как правило, с установления эмпирического, или частотного

распределения, которое можно считать моделью вероятностной системы в первом приближении. Частотное распределение может дать достаточно хорошее представление об устойчивых чертах в структуре и функционировании исследуемой системы. Моделирование данных с помощью какого-либо теоретического распределения поднимает исследование на более высокую ступень обобщения. Известна важная роль теоретических распределений в отображении закономерностей материального мира, и можно предположить, что в отношении некоторых классов лингвистических объектов такое моделирование открывает широкие возможности не только для решения многих актуальных прикладных задач, но и для выявления более глубоких квантитативных закономерностей структуры и функционирования языка.

В самом процессе моделирования направление мысли идет в начале от объекта к модели (формирование модели), а затем от модели обратно к объекту (интерпретация модели, получение нового знания об объекте).

Интерпретация лингвистических распределений. По своей природе лингвистические распределения не являются чисто формальными моделями, их надо рассматривать как интерпретируемые знаковые системы. Выявление свойств и отношений, а также внутренних закономерностей, которые определяют характер лингвистического распределения в целом, должно сопровождаться качественным (содержательным) анализом результатов исследования.

Интерпретацию (объяснение, толкование, раскрытие смысла) лингвистических распределений как моделей изучаемых вероятностных систем можно представить себе как многоуровневый (многостадиальный) процесс качественно-количественного анализа и синтеза. Задача интерпретации состоит в том, чтобы раскрыть сущность явления, однако здесь могут быть разные пути к достижению поставленной цели. Различаются индуктивный и дедуктивный подходы, причем высшей формой интерпретации считается научное объяснение, которое должно показать, что "данный научный факт является проявлением определенного закона или что объясняемый закон вытекает из более общего закона (или же теории)" (Друянов Л.А., 1980, с. 53). В зависимости от условий и задач исследования и по характеру материала при исследовании лингвистических распределений можно условно выделить три типа интерпретаций: структурно-функциональную, прагматическую (стилистическую) и генетическую (причинную). На практике эти три типа интерпретаций, будучи вза-

имосвязанными и взаимопроникаемыми, могут сосуществовать в одном конкретном исследовании.

I. Структурно-функциональная интерпретация

Важным условием моделирования является проведение предмодельного анализа физической сущности изучаемого явления с целью формирования некоторой априорной информации для вывода общего вида искомой модели. Наиболее простым и удобным средством представления информации о распределении является г р а ф и к о - г е о м е т р и ч е с к и й м е т о д. Графики (рисунки) позволяют представить данные в наглядной форме при минимальной их обработке. По графическому изображению можно сделать обоснованные выводы об общей форме распределения и о характере взаимосвязи элементов изучаемой системы (симметричность - несимметричность, линейность - нелинейность, унимодальность - мультимодальность и т.д.). S-образная кривая говорит о том, что данное распределение может отражать динамический процесс, подчиняющийся логистическому закону роста, а гипербола на графике может указать на феномен "концентрации и рассеяния" элементов системы. Графики могут быть представлены в обыкновенной декартовой системе координат или в модифицированной форме, например, в логарифмическом масштабе, на котором легко можно удостовериться в соответствии или несоответствии данных определенным законам распределения. Важно отметить ценность графиков при попытках добыть новую информацию по отклонениям от общей тенденции в отдельных частях распределения, по точкам перегиба и т.п. Так, например, анализ формы кривой распределения частот слов в билогарифмических координатах дал толчок для выявления особого "нелинейного" типа рангового распределения лексики в больших текстах и для формулировки тезиса о "четвертом приближении закона Ципфа" (Алексеев П.М., 1978).

В отличие от графиков ф о р м у л ы обладают тем важным преимуществом, что допускают проведение различных операций, которые сами по себе могут служить основанием для раскрытия новых, неожиданных сторон изучаемого явления. Формулу можно рассматривать как символическую (аналитическую) запись структуры данного явления, но формула может дать также представление о функционировании системы, о динамическом процессе, о развитии и т.п.

В качестве аналитических записей распределений в квантитативной лингвистике обычно употребляются различные функ-

ции или дифференциальные уравнения. Представляя распределение в виде функции (функциональной зависимости)[+], часто начинают с определения вида функции по опытным данным. Такой индуктивно-эмпирический подход может иногда дать вполне приемлемые результаты. Например, если установлено, что данное эмпирическое распределение описывается степенной функцией типа $y = ax^b$ (где a и b - параметры), то можно сделать вывод об "аллометрическом" законе изменения y в зависимости от x, причем параметр b ("коэффициент относительной эластичности") позволяет определить средний процент изменения y в связи с изменением x на I %, а параметр a указывает на начальное значение y при $x = I$. Безусловно, качественная интерпретация формулы и ее параметров требует анализа явления или процесса по существу, по внутренней логике или по физической сущности с целью формирования адекватного представления об изучаемом явлении. Далее, представляя эту же формулу в дифференциальной записи, например, в форме уравнения $\frac{dy/y}{dx/x} = b$, мы видим, что соотношение относительных приростов y и x постоянно (сохраняет устойчивость) и тем самым явление подчиняется закону "постоянного относительного роста" (Ланд К.Ч., 1977, с. 388) - одному из важнейших законов, характеризующих некоторые типы сложных самоорганизующихся систем.

Аналогично может интерпретироваться экспоненциальная функция как "лавинообразный рост", логарифмическая функция как "закон адаптационного торможения" (Налимов В.В., Мульченко З.М., 1969, с. 41) или "закон пропорционально убывающего относительного роста" (Ланд К.Ч., 1977, с. 388), логистическая функция как "рост с начальным ускорением и последующим замедлением (и насыщением)", функция Вейбулла как "обобщенная модель прогрессивного роста (включающая экспоненциальный закон как частный случай)" (см. Добров Г.М., 1969, с. 158) и т.д.

На другом (дедуктивном) уровне научного анализа используются гипотетико-теоретические модели вероятностных систем, т.е. когда те или другие функции-модели выводятся гипотетически, на основе некоторых теорети-

---

[+] В отношении вероятностных систем функциональная зависимость трактуется как вероятностная функция распределения. Формально между ними нет разницы, пока мы не вкладываем определенного смысла в математические символы.

ческих постулатов большей или меньшей степени общности. Интерпретация таких моделей содержит в себе как сами исходные постулаты, так и возможные выводы, сделанные на основе анализа конкретного материала. В квантитативной лингвистике известны попытки вывода закона Ципфа, используя понятие вероятностного процесса (например, Simon H.A., 1955) или прибегая к аналогии с термодинамикой (Mandelbrot B., 1954). Некоторым авторам удалось вывести формулы квантитативной структуры текста, исходя из комбинаторно-вариационных принципов (Арапов М.В., Шрейдер Ю.А., 1978; Крылов Ю.К., 1987). Модель "обобщенного закона Ципфа-Мандельброта" (Орлов Ю.К., 1976) выводится теоретически, но тезис об особой роли "объема Ципфа" обосновывается и интерпретируется, исходя из эмпирических соображений, причем делается вывод, что выполнение закона Ципфа (соответствие "объему Ципфа") свидетельствует о "высокой степени организованности целостного текста".

В квантитативной лингвистике практикуется также промежуточный "г и п о т е т и к о - э м п и р и ч е с к и й" подход, при котором исходная модель получается на основе некоторых относительно самостоятельных теоретических схем (гипотез), а конкретную форму распределения определяют эмпирически итеративным путем, придерживаясь требований исходной теоретической модели (см. гл. 2.3).

Далее, имеются интересные попытки сконструировать и интерпретировать теоретические модели, исходя из анализа в з а и м н ы х з а в и с и м о с т е й ("синергетики") между лингвистическими объектами (Altmann G., 1980; Köhler R. 1986). Конечной целью в этих исследованиях объявляется построение адекватной лингвистической теории в рамках общей теории саморегулирующихся систем.

2. Прагматическая (стилистическая) интерпретация

Под этим названием подразумевается объяснение, которое основывается на связи формальных (квантитативных) показателей с некоторыми прагматико-стилистическими, в т.ч. оценочными характеристиками изучаемых явлений. В квантитативной лингвистике, в частности в ее подразделении - квантитативной лингвостилистике, или стилеметрии, оперируют такими содержательными (стилистическими) понятиями как "богатство" или "разнообразие" словаря, "целостность" и "художественная завершенность" текста, "стереотипность" и "экономия" высказывания и т.п. Все подобные понятия должны по замыслу исследователей охарактеризовать некоторые реальные свойства ре-

альных объектов, однако эти свойства прямо не наблюдаемы и тем более прямо не измеряемы. Тогда обращаются к внешним наблюдаемым сторонам изучаемых реальных объектов (словарей, текстов) и пытаются таким путем косвенно добраться до сути изучаемых явлений.

Теоретически такую ситуацию можно сравнивать с положением, известным из теории измерения, где рассматриваются переменные двух типов: I) л а т е н т н ы е (скрытые) п е р е м е н н ы е, т.е. то, что исследователь выбирает, фиксирует для своего анализа; 2) и н д и к а т о р ы - то, что непосредственно измеряется (Хайтун С.Д., 1983, с. 16). При этом необходимым условием анализа является существование определенной связи между латентными переменными и индикаторами.

К формальным индикаторам можно отнести и лингвистические распределения. Известно, например, что "богатство" словаря определяется по особенностям формы рангового распределения слов (в частности по углу наклона кривой или по ципфовскому параметру γ) или по комплексному распределению, отражающему связь между ростом объема словаря и ростом объема текста (по этой связи можно прогнозировать тенденцию роста, т.е. "потенциальное богатство" словаря). Асимметричное распределение частот слов истолковывается как осуществление принципа "предпочтения" или "значимости" определенной части словаря в данных условиях, что приводит к "концентрации и рассеянию" единиц. "Действенность" и "качественность" стиля определяется по распределению частей речи в тексте. И т.д.

При таком анализе необходимо помнить, что связь квантитативных индикаторов с латентными переменными (свойствами, характеристиками) носит вероятностный характер. Это значит, во-первых, что интерпретация данных может дать положительный эффект лишь на представительном материале, т.е. при достаточно больших выборках и при воспроизводимости результатов испытаний. Во-вторых, надо отдать себе отчет в том, что формальные индикаторы могут лишь косвенно отражать содержательно-латентные свойства изучаемых явлений. Возникает также вопрос, в какой мере выбранные индикаторы способны охватывать все существенные стороны явления и не может ли случиться искажение ("деформация") картины при неправильном выборе или недостаточной представительности индикаторов. Оправдано ли соотнесение данного индикатора (или данных индикаторов) с данной латентной переменной, в каждом конкретном случае про-

веряется практикой.

3. Генетическая (причинная) интерпретация

Генетической можно назвать такую интерпретацию, которая пытается объяснить изучаемое явление, указывая на его происхождение и тем самым прямо или косвенно устанавливая причину его существования. Здесь, так же как и при предыдущих типах интерпретации, надо иметь в виду, что мы имеем дело с вероятностным подходом к исследованию языковых явлений. При таком подходе надо исходить из вероятностной концепции причинности, согласно которой "причинность есть нечто, могущее присутствовать в большей или меньшей степени, а не только быть или не быть" (Винер Н., 1964, с. 309). Надо также признать принцип множественности причин, обусловливаемой множественностью связей данного явления с другими явлениями.

При исследовании лингвистических распределений можно, конечно, кое-что объяснить "внутренними", л и н г в и с т и ч е с к и м и причинами, например, особенностями морфологической структуры данного языка. Однако такие внутренние причины всегда связаны и переплетаются с внешними (внелингвистическими) причинами (контакты языков, общественные потребности и т.п.). В последнее время особый интерес представляют попытки психологического (психолингвистического), психофизиологического и филогенетического объяснения лингвистических распределений как моделей определенных сторон речевой деятельности.

Так, например, "гиперболическое" распределение частот слов (в виде степенной зависимости переменных), известное под названием закона Ципфа, пытаются связать с особенностями п с и х и к и ч е л о в е к а, его потребностью общения, с одной стороны, и стремлением свести к минимуму свои умственные и физические усилия, с другой стороны. Этот широко известный принцип "наименьшего усилия" (Zipf G.K., 1949) основывается, таким образом, на взаимодействии двух противоборствующих тенденций в психике (подсознании) человека. Взаимодействием противоположных тенденций при порождении речи (разнообразие - ограничение разнообразия) и ассоциативными свойствами человеческой памяти (при ограниченности объема кратковременной памяти) объясняются и некоторые другие известные лингвистические распределения, например, комплексное распределение, отражающее более медленный темп роста объема словаря по сравнению с темпом роста объема текста (см. гл. 2.3).

Представителями психофизиологии некоторые типы лингвистических распределений как моделей речевой деятельности связываются со структурными особенностями мозга, в частности с пространственно-временной организацией периодических (циклических) п р о ц е с с о в в м о з г е. Исходя из представления о кодировании образов слов "пакетами волн нейронной активности", А.Н. Лебедев выводит формулу, совпадающую с формулой Ципфа, для описания распределения частот слов в речи, и другую формулу, описывающую связь между ростом объема словаря и ростом объема текста (Лебедев А.Н., 1983 и 1986). Можно констатировать, что предположения о связи между особенностями квантитативной структуры текста и некоторыми закономерностями деятельности мозга, по-видимому, не лишены основания, и исследования в этой области приобретают актуальность.

В филогенетическом плане лингвистические распределения объясняются э в о л ю ц и о н н ы м р а з в и т и е м человеческого языка, который формировался на протяжении тысячелетий в результате приспособления к внешнему миру. Этот эволюционный процесс развития языка можно "отдаленно уподобить органической эволюции на основе естественного отбора" (Панов Е.Н., 1980, с. 147). Такое системное свойство основных лингвистических распределений как иерархичность можно объяснить адаптацией к увеличению числа элементов (при порождении речи), учитывая, что иерархическая структура минимизирует число связей (см. Козачков Л.С., 1978, с. 15). Устойчивость некоторых лингвистических распределений (сохранение общей формы) может указать на стремление к равновесию, оптимальности и целесообразности самоорганизующейся сложной системы - языка.

Безусловно, многие из названных свойств имеют всеобщий характер, они обнаруживаются в живой и неживой природе. Это свидетельствует о "единстве мира", которое состоит, в частности, в том, что "более общие законы "ниже" лежащих уровней бытия сохраняют свою силу для всех "выше" лежащих уровней" (Брушлинский А.В., 1979, с. 47), причем эта универсальность "не только исключает, а, наоборот, предполагает наличие специфических закономерностей" (там же).

В итоге можно констатировать, что генетический метод объяснения, ставящий на первое место отыскание причин явлений, так же как и прагматический метод, основанный на анализе латентных переменных, занимают прочное место среди методов интерпретации лингвистических распределений.

2. СТАТИСТИЧЕСКАЯ ОРГАНИЗАЦИЯ СЛОВАРЯ И ТЕКСТА

В главе рассматриваются вопросы составления и анализа частотных словарей и обсуждаются основные закономерности статистической организации словаря и текста с точки зрения квантитативно-системного подхода к изучению лексики (соотношение словоформ и лексем в тексте, стратификация слов по частотности, частотная структура текста, зависимость "словарь - текст").

2.1. ЧАСТОТНЫЕ СЛОВАРИ

О составлении частотных словарей. Одной из важнейших задач квантитативной лингвистики является составление частотных словарей, эффективность использования которых для решения разных прикладных и исследовательских задач постоянно возрастает. Частотный словарь дает представление о статистической структуре словаря, а также того текста, который послужил основанием для составления частотного словаря. Частотный словарь можно рассматривать как своего рода модель особым образом преобразованного текста, как м о д е л ь р а с п р е д е л е н и я ч а с т о т употребления единиц в тексте.

Частотный словарь (ЧС) представляет собой упорядоченный список слов, снабженных данными о частоте их употребления в тексте (речи). По способу размещения единиц различаются два основных типа ЧС:

1) алфавитно-частотный словарь (АЧС);

2) ранговый частотный словарь (РЧС).

В первом случае (в АЧС) слова расположены в алфавитном порядке с указанием частотности каждого слова. Во втором случае (в РЧС) слова расположены в порядке убывающей частотности с указанием (или без указания) ранга и частотности слова.

Особую разновидность АЧС представляют обратно-частотные словари (ОЧС), в которых слова (с указанием их частотности) расположены по алфавиту, начиная с конца слова.

Единицами ЧС могут выступать словоформы или лексемы. Для специальных целей составляются ЧС, в которых единицами являются основы слов, лексико-семантические варианты слов, сло-

восочетания и др. Материалом для составления ЧС может служить как индивидуальный текст, так и набор текстов, причем в обоих случаях за основу берутся либо целостные (законченные) тексты, либо отрывки (выборки) из текстов. (Подробнее о типологии и применении частотных словарей см. Алексеев П.М., 1975 и 1980; Alekseev P., 1984).

При составлении ЧС и при последующем анализе данных следует учитывать некоторые общие требования квантитативной лингвистики. Важное значение имеет вопрос о достоверности (надежности) материала. Этот вопрос подробно освещен в литературе. Отметим только, что в практике составления ЧС прежде всего необходимо уделять внимание о т б о р у и д о з и р о в к е лингвистического материала. В основу формирования выборочной совокупности текстов принимается обычно "минимальная", или "стандартная" выборка – отрезок текста длиной в 1000 словоупотреблений (Андреев Н.Д., 1967; Алексеев П.М., 1968; Якубайтис Т.А., 1981). Утверждается, что такая минимальная выборка обладает всеми чертами, присущими "связному" тексту (Якубайтис Т.А., Скляревич А.Н., 1978, с. 62) и поэтому подходит для решения самых разнообразных задач квантитативной лингвистики. Важно учитывать принцип однородности материала, поэтому обычно ограничиваются исследованием одного какого-нибудь подъязыка или только индивидуальных текстов и их сравнением. ЧС "всего языка" – это условное понятие, так как квантитативные характеристики такого словаря во многом зависят от состава и процентного распределения текстов подъязыков, легших в основу составления словаря. На практике в вопросах дозировки и определения объема большого ЧС наблюдаются большие расхождения. Например, ЧС русского языка (1977) составлен на основе текстов общим объемом 1 милл. словоупотребдений (СУ), которые взяты примерно в равных пропорциях из четырех жанровых групп текстов: художественная проза (25,4 %), драматургия (27,2 %), научно-публицистические тексты (23,6 %) и газетно-журнальные тексты (23,8 %). ЧС словацкого языка (Mistrík J., 1969) составлен на основе текстов объемом 1 милл. СУ по следующему плану: художественная проза (30,2 %), научно-технические тексты (31,5 %), газетно-журнальные тексты (14,6 %), поэзия (13,2 %), "диалог" (10,5 %). ЧС французского языка (Juilland A. et al., 1970) имеет своей основой тексты 5 жанров по 100 тыс. СУ: художественная проза, деловая проза, публицистика, драма, эссе. В основу сводного ЧС

финского языка положены тексты общим объемом 400 тыс. СУ, которые распределяются следующим образом: художественная проза (II,5 %), радиопередачи (I9,2 %), публицистика (26,0 %), разное (43,3 %). Общий объем текстов в 400 тыс. СУ является основой для составления многих других словарей, например, ЧС испанского языка, составленный на основе текстов четырех подъязыков в равных пропорциях (García Hoz, I953).

По вопросу об объеме текста подъязыка существуют разные мнения. Обычно считается, что для составления ЧС подъязыка требуется совокупность текстов (выборок) общим объемом не менее 200 тыс. СУ (при микровыборках по I000 - I0 000 СУ). Наш опыт показывает, что допустимым минимумом для одного подъязыка можно считать объем в I00 тыс. СУ (при микровыборках по I 000 - 5 000 СУ) при условии, что набор текстов такого объема дает не менее 2000 слов (лексем) с частотой $F \geqslant 5$, которые покрывают в тексте около 80 % (Tuldava J., I977, с. I49). Частота $F = 5$ имеет теоретические доверительные границы I....9, т.е. с нижней границей более 0, если предположить, что редкие слова распределяются приблизительно по Пуассону (ошибка вычисляется по упрощенной формуле $2\sqrt{F}$ при доверительном уровне 95 %). Такие статистические характеристики являются достаточными, чтобы выявить ядро лексики данного подъязыка, причем ЧС должен быть составлен по принципу распределительного словаря, позволяющего судить о с т а б и л ь н о с т и частот слов по отдельным подвыборкам. При надобности можно, например, исключить из рассмотрения слова, появляющиеся только в I-2 подвыборках. По данным распределительного словаря можно вычислить коэффициенты стабильности и употребительности слов (Juilland A., I970; Андрющенко Б.М., I978) или использовать другие способы оценки д о с т о в е р н о с т и частот с учетом действительного распределения слов по подвыборкам (например, Перебейнос В.И., I984). Окончательное решение вопроса о достаточном объеме текста для составления частотного словаря зависит от целей и задач исследования.

Таким образом, если качественная надежность ЧС обусловливается подбором текстов, то количественная, т.е. статистическая надежность ЧС оценивается, исходя из специальных требований, выработанных в современной квантитативной лингвистике. Притом следует отметить, что представление о ЧС как о списке слов, расположенных в строго частотном порядке, едва ли оправдывается с практической и теоретической

точек зрения. Для решения большинства актуальных квантитативно-лингвистических задач, в частности задач изучения закономерностей статистической организации лексики, нет необходимости получения "абсолютного частотного словаря". Отмечается, что смысл составления ЧС заключается прежде всего в "стратификации разных по статистическому весу пластов лексики" (Засорина Л.Н., 1966, с. 70), т.е. в выявлении основных частотных зон слов. Учитывая специфику квантитативного изучения лингвистических объектов, некоторые исследователи считают, что нет смысла говорить о точных вероятностях употребления слова в тексте, а в качестве инварианта употребления устанавливают "степень размытости" данного слова, т.е. спектр допустимых частот или рангов слова в определенного рода текстах (Арапов М.В. и др., 1978). С точки зрения квантитативно-системного подхода упорядоченность слов по частоте употребления рассматривается в плане вероятностных систем, т.е. в плане взаимодействия устойчивости и вариативности, причем основной упор делается на изучение системных свойств лексики, в частности, на изучение взаимосвязей и группировок слов и на моделирование с помощью разного рода распределений.

Основные характеристики ЧС. Наиболее общими статистическими характеристиками ЧС являются: N – объем текста, на основе которого составлен словарь (т.е. число словоупотреблений в тексте); V – объем словаря словоформ; L – объем словаря лексем; а также относительные показатели: V/N (или L/N) – отношение объема словаря к объему текста (показывает относительное "богатство", или "разнообразие" словаря); обратное соотношение N/V (или N/L) выражает среднюю частоту (повторяемость) слова в данном тексте.

При использовании упомянутых относительных статистических показателей следует учесть, что они зависят от объема, а также от типа текста.

Для примера приводим данные ЧС авторской речи современной эстонской художественной прозы:

	N	V	L	V/N	N/V	L/N	N/L
Подвыборка	5000	2690	1953	0,54	1,9	0,39	2,6
Сводный текст	99898	30733	14654	0,31	3,3	0,15	6,8

Можно констатировать, что при увеличении объема текста относительное разнообразие лексики (отношение V/N или L/N)

Таблица 2.I

Основные статистические характеристики ЧС разных языков (Алексеев П.М., I968; Бектаев К.Б., I978; Григорьева А.С., I98I; Част. слов. совр. укр. худ. прозы (пробн. тетрадь), I969; Kučera H., Francis W.N., 1967; Latviešu val. biež. vārdn., 1972)

Язык	N	V	L	V/N	N/V	L/N	N/L
Эстонский (худ. проза)	99 898	30 733	I4 654	0,3I	3,3	0,I5	6,8
Украинский (худ. проза)	I00 000	27 570	I3 954	0,28	3,6	0,I4	7,2
Латышский (худ. проза)	I00 000	-	II 439	-	-	0,II	8,7
Казахский (детск. лит.)	98 040	23 350	I0 076	0,24	4,2	0,I0	9,7
Русский (эпист. речь)	96 800	I5 842	8 064	0,I6	6,I	0,08	I2,0
Немецкий (научно-техн.)	I00 000	I4 434	-	0,I4	6,9	-	-
Английский (общелит. тексты)	I0I 566	I3 706	-	0,I3	7,4	-	-
Английский (электроника)	I00 000	7 853	5 I97	0,079	I2,7	0,052	I9,2
Французский (электроника)	I00 000	8 I08	4 572	0,08I	I2,3	0,046	2I,9

уменьшается, а средняя повторяемость слов (N/V или N/L) увеличивается. Если сравнивать сводный текст авторской речи художественной прозы (состоящий из 20 подвыборок по ~ 5000 СУ из текстов разных авторов) с индивидуальным текстом авторской речи (по данным романа А.Х. Таммсааре; см. Villup A., 1978), то оказывается, что в индивидуальном тексте значения всех относительных показателей значительно отличаются от соответствующих показателей сводного текста примерно такого же объема:

Роман	N	V	L	V/N	N/V	L/N	N/L
А.Х. Таммсааре	II4124	I6750	7348	0,I5	6,8	0,64	I5,5

Особенно бросается в глаза разница в значениях N/V (и N/L): в сводном тексте (в наборе выборок из текстов разных авторов) средняя повторяемость слов меньше и, следовательно, разнообразие лексики больше, чем в индивидуальном целостном тексте такого же объема.

При сравнении данных разных языков (см. табл. 2.I) отчетливо выявляется т и п о л о г и ч е с к о е различие языков. Если, например, в синтетическом эстонском тексте (объемом около I00 тыс. СУ) словоформа повторяется в среднем 3,3 раза, то в аналитическом английском тексте (такого же объема) средняя частота словоформы колеблется от 7,4 (в общелитературном тексте) до I2,7 (в научно-техническом тексте). При функциональной однородности текстов и типологической близости языков средняя повторяемость словоформы оказывается близкой, например, в текстах электроники аналитических английского и французского языков (N/V равно I2,7 и I2,3, соответственно). Влияние функционального стиля сказывается при сравнении данных синтетических языков – эстонского (художественные тексты – N/V = 3,3) и казахского (детская литература – N/V = 4,2).

Соотношение словоформ и лексем. Важным квантитативно-типологическим показателем языка является количественное соотношение L/V (или V/L), т.е. отношение числа разных лексем к числу разных словоформ (или наоборот) в данном тексте. Это соотношение характеризует в известной мере морфологическую структуру языка с квантитативной точки зрения и позволяет судить о степени а н а л и т и з м а / с и н т е т и з м а языка. Чем больше численное отношение L/V, тем более аналитичным является язык данного текста, так как в этом случае число разных лексем приближается к числу разных словоформ и, следовательно, в тексте появляется в среднем мень-

ше словоизменительных форм на каждую лексему. Наоборот, большее численное отношение V/L говорит о том, что в тексте на каждую лексему приходится в среднем больше словоизменительных форм, и язык такого текста следует признать более синтетичным. Однако при этом надо учесть, что количественные меры аналитизма/синтетизма чувствительны к изменению объема текста (N). Опыт показывает, что по мере увеличения N (до определенных пределов) отношение L/V неустанно уменьшается, а отношение V/L соответственно увеличивается. Такую тенденцию можно лучше всего проиллюстрировать на материале индивидуального текста (см. табл. 2.2). Однако при увеличении текста до больших размеров все больше и больше будет сказываться влияние определенной закономерности появления редких слов: при переходе к выборкам большего объема практически все появляющиеся заново слова - редкие (ср. Фрумкина Р.М., 1964). Тем самым замедляется темп уменьшения отношения L/V (и увеличения отношения V/L, соответственно).

Сравнение текстов разных языков (при соблюдении одинаковых условий эксперимента) показывает, что отношение L/V (или V/L) позволяет отграничить языки более аналитичные от менее аналитичных (табл. 2.3). Но при этом оказывается, что меры аналитизма/синтетизма зависят не только от объема текста и от языка, но в определенной степени также от функционального стиля. Например, при объеме текста N = 200 000 в английском языке коэффициент аналитизма (отношение L/V) для газетного текста равняется 0,53, а для научно-технического текста - 0,67. Можно сделать вывод, что основное влияние на численную меру аналитизма оказывает объем словаря: в газетном тексте объем словаря $L \approx$ 12 000, в то время как в научно-техническом тексте такого же объема количество разных лексем лишь около 7000. Ясно, что соотношение числа лексем к числу словоформ при такой разнице объемов словарей изменяется, а зависимость от объема текста проявляется лишь косвенно, т.е. в той мере, в какой объем словаря зависит от объема определенного типа текста.

Ввиду того, что численная мера аналитизма/синтетизма в первую очередь определяется размером словаря данного текста, целесообразно выявить аналитическую зависимость между количеством словоформ (V) и количеством лексем (L) в связи с изменением одной из этих величин. Этот вопрос имеет не только теоретическое, но и практическое значение. В частности при рассмотрении словарей, составленных на основе текстов

Таблица 2.2

Динамика изменения мер аналитизма/синтетизма по данным авторской речи I-го тома романа А.Х. Таммсааре "Правда и право" ("Tõde ja õigus") - эст. яз.

N	V	L	L/V	V/L
I0 000	3636	2II4	0,58	I,72
20 000	5944	3I24	0,53	I,90
30 000	7503	378I	0,50	I,98
II4 I24	I6750	7348	0,44	2,28

Таблица 2.3

Меры аналитизма/синтетизма разных языков (данные взяты из работ: Алексеев П.М., I975; Бектаев К.Б., I978; Яблонская Н.Н., I976; Engwall G., 1974)

Язык	N	V	L	L/V	V/L
Русский (электроника)	200 000	2I 648	6 8I6	0,32	3,I8
Румынский (электроника)	200 000	I4 292	5 708	0,40	2,50
Немецкий (медицина)	200 000	4I 04I	20 367	0,50	2,02
Французский (худ. лит.)	200 000	20 53I	I0 868	0,53	I,89
Английский (газеты)	200 000	23 595	I2 588	0,53	I,87
Английский (электроника)	200 000	I0 582	7 I60	0,67	I,48

синтетических языков, может возникнуть необходимость прогнозировать объем словаря лексем, зная объём словаря словоформ, или наоборот.

Мы исходим из предположения, что в определенных пределах (исключая тексты очень большого объема) в процессе порождения речи (текста) существует постоянное отношение между скоростью относительного роста словаря лексем и скоростью относительного роста словаря словоформ. Это вполне естественное предположение, которое соответствует закону "аллометрического", или "постоянного относительного роста". Математически закон выражается в форме дифференциального уравнения:

$$\frac{dy/y}{dx/x} = b. \qquad (2.1)$$

Формула может быть переписана в следующей форме:

$$dy/y = b\,(dx/x). \qquad (2.2)$$

Интегрируя, получаем

$$\ln y = A + b \ln x, \qquad (2.3)$$

что показывает линейную связь между $\ln y$ и $\ln x$. Если взять $A = \ln a$, мы получим степенную функцию (аллометрическую функцию роста)

$$y = ax^b, \qquad (2.4)$$

где a и b - параметры.

Принимая $y = V$ и $x = L$, мы можем представить формулу для выражения зависимости между числом разных словоформ и числом разных лексем в данном тексте в следующем виде:

$$L = aV^b, \qquad (2.5)$$

где a и b - параметры.

Проверка показывает, что формула (2.5) дает хорошее соответствие между эмпирическими (наблюдаемыми) и теоретическими (ожидаемыми) величинами по данным эстонского, русского и некоторых других языков (см. также Тулдава Ю.А., в печати; о других подходах к анализу соотношения словоформ и лексем см. Орлов Ю.К., 1978; Нешитой В.В., 1975). При условии однородности материала можно с успехом прогнозировать число лексем (L) при заданном числе словоформ (V) по предложенной формуле (2.5). Например, на основе двух первых выборок из романа А.Х. Таммсааре (табл. 2.2):

N_1 = 10 000 V_1 = 3636 L_1 = 2114

N_2 = 20 000 V_2 = 5944 L_2 = 3124

вычисление параметров дает результат: $a = 3{,}1$ и $b = 0{,}8$.[+] Прогноз на весь роман (при N = 114 124 и V = 16 750) по формуле (2.5): $L = 3{,}1 \cdot 16\,750^{0{,}8} = 7423$, т.е. хорошее приближение к наблюдаемому числу лексем (L = 7348).

В формуле (2.5) параметр b имеет ясный содержательный смысл и раскрывает истинное соотношение между V и L: например, если $b = 0{,}8$, то это значит, что если V увеличивается на 1 %, то L увеличивается в среднем на 0,8 %. В этом смысле параметр b может быть интерпретирован как показатель аналитизма/синтетизма языка: очевидно, что при большем значении b язык данного текста более аналитичен (при увеличении числа разных словоформ число лексем растет быстрее). Надо иметь в виду, что при сравнении языков или текстов должны соблюдаться одинаковые условия эксперимента, в частности важно то, как определяется лексема (считается ли лексемой совокупность всех лексико-семантических вариантов, или они считаются отдельными лексемами). Надо также считаться с влиянием функционального стиля. Предварительный опыт говорит о том, что в однородном тексте параметры a и b остаются довольно устойчивыми в пределах $500 < N < 200\,000$.

На практике может возникнуть необходимость установить обратное соотношение, т.е. зависимость V от L. В таком случае следует "перефразировать" формулу (2.5) таким образом, что получается

$$V = \alpha L^{\beta}, \qquad (2.6)$$

где $\alpha = e^{-(\ln a/b)}$ и $\beta = 1/b$.

Наконец, можно поставить вопрос о связи между показателем аналитизма/синтетизма, т.е. отношением L/V (или V/L) и объемом текста N. Между ними существует корреляция, хотя, как уже было отмечено, определяющую роль при формирова-

---

[+] Параметры a и b можно вычислить методом наименьших квадратов на основе линеаризации (на основе линейной связи между $\ln L$ и $\ln V$). В данном случае – при наличии двух точек – параметры вычисляются просто:

$$b = (\ln L_1 - \ln L_2)/(\ln V_1 - \ln V_2);$$

$$a = e^{A}, \text{ где}$$

$$A = [(\ln L_1 + \ln L_2) - b(\ln V_1 + \ln V_2)]/2.$$

ния меры аналитизма/синтетизма играет объем словаря данного текста (при одинаковых объемах текстов объемы их словарей могут сильно варьироваться в зависимости от особенностей функционального или индивидуального стилей). Отношение L/V в среднем уменьшается в связи с увеличением N (см. табл. 2.2). Оказывается, что связь между этими величинами не линейная, а также носит характер степенной зависимости, которую можно выразить формулой

$$L/V = c\,N^{-d} \qquad (2.7)$$

где c и d - параметры.

Таким образом, знание закономерности динамики соотношения "словоформа - лексема" позволяет прогнозировать взаимно V и L или в определенных пределах вычислять численную меру аналитизма/синтетизма для разных значений N. Все это может иметь значение для типологического исследования языков и для решения некоторых задач автоматической переработки текста.

Распределения и частотные зоны словаря. Важное значение для квантитативно-лингвистического исследования лексики имеет выяснение типа (закона) распределения частот слов по "горизонтали", т.е. по подвыборкам (микровыборкам) из данной совокупности ("однообъектные" распределения; см. гл. I.3). Вопрос о такого рода распределениях многократно обсуждался в литературе (см., например, Herdan G., 1964; Бектаев К.Б., Лукьяненков К.Ф., 1971; Пиотровский Р.Г., Турыгина К.Ф., 1971; Камирина М.Е., 1974; Якубайтис Т.А., 1981). Выявилось, в частности, что на основе определения типа распределения можно отличить семантически доминантные (ключевые, терминологические) единицы текста от нейтральных, общеупотребительных единиц ("эффект Бектаева"). Но при этом следует иметь в виду, что определение типа распределения во многом зависит от условий исследования (например, от величины подвыборок); изменение условий исследования может привести к замене одного типа распределения другим (Бектаев К.Б., Лукьяненков К.Ф., 1971, с. 104). По данным некоторых исследований большинство общеупотребительных высокочастотных и среднечастотных слов подчиняется нормальному или пуассоновскому распределению при микровыборках объемом не менее 2 тыс. словоупотреблений. Распределение редких лексических единиц независимо от разбивки и нормировки выборки подчиняется закону Пуассона (Пиотровский Р.Г., 1975; Piotrowski R.G., 1984).

Факт подчинения распределения частот слов нормальному или пуассоновскому закону говорит о том, что текст, на котором проверяется это распределение, является статистически однородным по отношению к этим словам. Однако это не исключает возможности выявления индивидуальных различий во множестве частот. В частности можно выявить т.н. плюсовые и минусовые слова при сравнении индивидуальных стилей на основе анализа отклонений от средних частот (см., например, Muller Ch., 1968, с. 87; McKinnon A., 1980).

Как уже было упомянуто выше, одной из важнейших задач составления ЧС является выявление ч а с т о т н ы х з о н словаря. Стратификация лексики по признаку частотности имеет большое значение не только для научных исследований в области лингвистики (лексикологии и лингвостилистики), но и для педагогики и психологии (составление "словарей-минимумов", измерение трудности учебных текстов, проведение психологических опытов), а также для решения проблем автоматической переработки текста (в частности для рационализации поиска слов в автоматическом словаре). Важно отметить то, что количественные характеристики слов, относящиеся к какой-нибудь частотной зоне, во многом коррелируют с качественными свойствами этих слов, например, с такими свойствами как нейтральность (общеупотребительность), тематичность ("ключевые слова"), информативность и др.

Представляют интерес многочисленные попытки формально-статистического разбиения словаря на отдельные частотные зоны. Чаще всего основанием для выделения зон служат особенности распределения частот слов. Например, частотные зоны выделяют в зависимости от колебания параметров закона Ципфа при ранговом распределении слов (Горькова В.И.,1969), в зависимости от точки "максимальной кривизны" интегрального распределения Вейбулла (Петренко Б.В., 1974) или по признаку изменения типа рангового распределения слов (Мартыненко Г.Я., 1978). Имеется еще ряд других способов формального разбиения ЧС на зоны (см., например, Маршакова И.В., 1974; Малаховский Л.В., 1980; Billmeier G., 1968; Rao M.L., 1978). Очевидно, что выделение частотных зон возможно по разным основаниям, и оно может быть осуществлено в зависимости от целей и задач конкретного исследования.

2.2. ЧАСТОТНАЯ СТРУКТУРА ТЕКСТА

Понятие частотной структуры. Если абстрагироваться от конкретных лексических единиц, составляющих частотный словарь, и рассматривать лишь частоты (F_i) и ранги (i) лексических единиц, то получается т.наз. ранговое распределение частот, или распределение "ранг-частота". Другой возможностью формального анализа ЧС является сопоставление частот F_i с численностью (количеством) единиц, имеющих данную частоту - $m(F_i)$, что дает спектральное распределение частот, или частотный спектр лексики. Совместное рассмотрение рангового и спектрального распределений частот лексических единиц раскрывает нам ч а с т о т н у ю с т р у к т у р у лексики текста, представляющую собой определенный аспект общей статистической организации текста (охватывающей всю проблематику анализа структуры и функционирования текста и соответствующего словаря в квантитативном освещении).

Определяемую таким образом частотную структуру текста можно представить наглядно в следующей компактной форме:

i	F_i	$m(F_i)$
1	F_1	$m(F_1)$
2	F_2	$m(F_2)$
...	...	...
$k-n$	F_{k-n}	$m(F_{k-n})$
...	...	...

На материале ЧС лексем авторской речи современной эстонской художественной прозы (Kaasik Ü. et al., 1977) комплексное распределение рангов, частот и численностей слов с данной частотой по указанной схеме приводится в табл. 2.4. Наряду с комплексом ранжированных величин исходными данными при рассмотрении частотной структуры текста являются также объем текста (N) и объем соответствующего словаря (L). В данном конкретном случае N = 99 898 (словоупотреблений) и $L = i_{max}$ = 14 654 (лексем).

Ранговое и спектральное распределения могут иметь дифференциальную (некумулятивную) форму, или они могут быть представлены в интегральной (кумулятивной) форме, т.е. в ви-

Таблица 2.4

Частотная структура текста по данным ЧС л е к с е м авторской речи эстонской художественной прозы (i - ранг, F_i - частота, $m(F_i)$ - количество слов с данной частотой). Объем текста N = 99 898; объем словаря L = 14 654; F_{max} = 4 237.

i	F_i	$m(F_i)$	i	F_i	$m(F_i)$	i	F_i	$m(F_i)$	i	F_i	$m(F_i)$	i	F_i	$m(F_i)$	i	F_i	$m(F_i)$
I	4237	I	41	264	I	80	I53	I	I32	96	I	239-24I	57	3	587-6I2	22	26
2	3493	I	42	26I	I	8I	I5I	I	I33-I35	94	3	242-244	56	3	6I3-650	2I	30
3	2598	I	43	260	I	82	I50	I	I36-I39	93	4	245-247	55	3	65I-676	20	26
4	I98I	I	44	254	I	83	I49	I	I40-I4I	92	2	248-252	54	5	677-707	I9	3I
5	I395	I	45	248	I	84	I48	I	I42-I44	9I	3	253-256	53	4	708-747	I8	40
6	I300	I	46	238	I	85	I46	I	I45-I47	90	3	257-263	52	7	748-793	I7	46
7	I047	I	47	237	I	86	I44	I	I48-I49	89	2	264-268	5I	5	794-856	I6	63
8	879	I	48	234	I	87	I39	I	I50	88	I	269-270	50	2	857-893	I5	37
9	845	I	49	230	I	88	I38	I	I5I	87	I	27I-274	49	4	894-940	I4	47
I0	827	I	50	223	I	89	I36	I	I52-I54	86	3	275-280	48	6	94I-I0I7	I3	77
II	724	I	5I	2I8	I	90	I35	I	I55-I57	85	3	28I-285	47	5	I0I8-I080	I2	63
I2	634	I	52	209	I	9I-92	I33	2	I58-I59	84	2	286-289	46	4	I08I-II69	II	89
I3	6I3	I	53	207	I	93-94	I3I	2	I60	83	I	290-299	45	I0	II70-I293	I0	I24
I4	58I	I	54	206	I	95-96	I30	2	I6I-I62	82	2	300-305	44	6	I294-I4I4	9	I2I
I5	568	I	55	200	I	97	I26	I	I63-I64	8I	2	306-3I2	43	7	I4I5-I567	8	I53
I6	499	I	56	I98	I	98-99	I25	2	I65-I69	79	5	3I3-3I8	42	6	I568-I779	7	2I2
I7	496	I	57	I95	I	I00	I24	I	I70-I72	78	3	3I9-329	4I	II	I780-2044	6	265
I8	493	I	58	I94	I	I0I	I23	I	I73-I79	76	7	330-339	40	I0	2045-2389	5	345
I9	465	I	59	I92	I	I02-I03	I22	2	I80-I83	75	4	340-346	39	7	2390-2980	4	59I
20	448	I	60	I90	I	I04	I20	I	I84-I87	74	4	347-349	38	3	298I-39I8	3	938
2I	436	I	6I	I89	I	I05	II6	I	I88-I9I	73	4	350-360	37	II	39I9-5972	2	2054
22	434	I	62-63	I85	2	I06	II3	I	I92-I94	72	3	36I-369	36	9	5973-I4654	I	8682
23	428	I	64	I8I	I	I07-I08	II2	2	I95-202	7I	8	370-382	35	I3			
24	382	I	65	I80	I	I09-II0	II0	2	203-204	70	2	383-389	34	7			
25	373	I	66	I79	I	III	I09	I	205-206	69	2	390-404	33	I5			
26	366	I	67	I76	I	II2	I08	I	207	68	I	405-4I2	32	8			
27	350	I	68	I75	I	II3-II5	I07	3	208-209	67	2	4I3-428	3I	I6			
28	339	I	69-70	I74	2	II6	I06	I	2I0-2II	66	2	429-445	30	I7			
29	327	I	7I	I7I	I	II7-II9	I05	3	2I2-2I7	64	6	446-466	29	2I			
30	309	I	72	I67	I	I20	I04	I	2I8-220	63	3	467-479	28	I3			
3I-33	304	3	73	I66	I	I2I-I24	I0I	4	22I	62	I	480-509	27	30			
34	297	I	74	I65	I	I25-I26	I00	2	222-223	6I	2	5I0-529	26	20			
35-37	285	3	75-76	I64	2	I27	99	I	224-229	60	6	530-543	25	I4			
38-39	272	2	77-78	I63	2	I28	98	I	230-235	59	6	544-564	24	2I			
40	268	I	79	I59	I	I29-I3I	97	3	236-238	58	3	565-586	23	22			

Таблица 2.5

Ранговое распределение частот по данным ЧС с л о в о - ф о р м авторской речи эстонской художественной прозы (i - ранг, F_i - частота, F_i^* - накопленная частота, P_i^* (%) - покрываемость текста); N = 99898; V = 30733; F_{max} = 3221.

i	F_i	F_i^*	P_i^*(%)	i	F_i	F_i^*	P_i^*(%)
I	3221	3221	3,22	200	48	35066	35,10
2	1602	4823	4,83	300	34	38592	38,63
3	1439	6262	6,27	400	25	41484	41,53
4	1375	7637	7,64	500	21	43755	43,80
5	1264	8901	8,91	600	17	45634	45,68
6	1116	10017	10,03	700	15	47264	47,31
7	995	11012	11,02	800	13	48703	48,75
8	713	11725	11,74	900	12	49966	50,02
9	592	12317	12,33	1000	11	51109	51,16
10	542	12859	12,87	1001-1034	11	51483	51,54
20	329	16801	16,82	1035-1168	10	52823	52,88
30	224	19344	19,36	1169-1299	9	54002	54,06
40	189	21384	21,41	1300-1500	8	55610	55,67
50	165	23130	23,15	1501-1753	7	57381	57,44
60	141	24635	24,66	1754-2131	6	59649	59,71
70	120	25920	25,95	2132-2650	5	62244	62,31
80	108	27045	27,07	2651-3460	4	65484	65,55
90	89	28017	28,05	3461-5088	3	70368	70,44
100	83	28871	28,90	5089-8973	2	78138	78,22
				8974-30733	1	99898	100,00

де накопленных частот или численностей.[+] Интегральное распределение выражает покрываемость текста, которая позволяет судить о степени концентрации лексических единиц в определенных участках ЧС. Естественно, что частоты и численности могут быть представлены как в абсолютных, так и относительных величинах (в т.ч. в процентах). Для примера приводим данные о ранговом распределении частот словоформ в сокращенной форме вместе с данными о покрываемости текста в абсолютных и относительных величинах (табл. 2.5).

Сначала мы рассмотрим одну из сторон частотной структуры - ранговое распределение - и возможности его аналитического описания с помощью разных вариантов закона Ципфа.

Ранговое распределение и закон Ципфа. Одной из важнейших закономерностей, выявленных при квантитативном анализе текстов, является статистическая связь между ч а с т о т о й и р а н г о м слова. При этом констатируется, что, хотя в различных текстах слова могут иметь различные ранги, все же устойчивой является сама форма распределения, т.е. вид закономерности в целом. Здесь сказывается "топологический" принцип, согласно которому "важна не "метрика" /.../, но важно сохранение "схемы", которая может видоизменяться, оставаясь самой собой и может "наполняться" разным содержанием" (Бернштейн Н.А., 1966, с. 65). Во всех случаях, когда мы имеем дело с текстами естественного языка, проявляется т.наз. эффект к о н ц е н т р а ц и и и р а с с е я н и я, который состоит в том, что имеется небольшая группа очень частых слов ("ядро" ЧС) и большая группа редких слов ("хвост" ЧС); между ними наблюдается плавный переход ("зона среднечастотных слов"). На графике это выражается формой, напоминающей гиперболу (см., например, рис. 2.1). Такая неравномерность в распределении единиц обнаруживается не только в отношении слов, но и в отношении других языковых единиц (букв, фонем, морфем, словосочетаний и т.д.). Более того, сходная форма распределения встречается и во многих других областях человеческой деятельности (информатике, экологии,

[+] Соответственно различаются дифференциальная функция (плотность) и интегральная функция распределения (см. Митропольский А.К., 1971, с. 209 и след.). В ином аспекте рассматриваются дифференциальное уравнение (включающее производные или дифференциалы) и интегрирование (т.е. решение дифференциального уравнения).

демографии и др.), что заставляет рассматривать этот тип распределения как универсальный семиологический "закон предпочтения" (Перебийніс В.С., 1970), или "закон распределения единиц по значимости" (Мартыненко Г.Я., 1978).

Для аналитического выражения зависимости между частотой и рангом слова предлагается множество формул, которые представляют собой разновидности закона Ципфа (Zipf G.K., 1935; 1949; Mandelbrot B., 1954; Орлов Ю.К., 1976; Алексеев П.М., 1978; Крылов Ю.К., 1982; и др.). Основная форма распределения Ципфа выражается следующей формулой, которая представляет собой степенную функцию с отрицательным показателем степени:

$$F_i = Ci^{-\gamma} \quad \text{или} \quad p_i = ki^{-\gamma}, \qquad (2.8)$$

где F_i - абсолютная частота, p_i - относительная частота (вероятность), i - ранг, C, k и γ - параметры распределения (причем $p_i = F_i/N$ и $C = kN$, где N - объем текста). В частном случае, когда $\gamma = 1$, формула принимает вид ("классическое однопараметрическое распределение Ципфа"):

$$F_i = Ci^{-1} \quad \text{или} \quad p_i = ki^{-1}. \qquad (2.9)$$

Каков содержательный смысл закона Ципфа? Чтобы раскрыть "механизм" изменения переменных, рассмотрим соответствующее функции (2.8) дифференциальное уравнение:

$$\frac{dF_i / F_i}{di / i} = -\gamma. \qquad (2.10)$$

Из уравнения явствует, что отношение скоростей относительного изменения F_i и i остается постоянным. Так характеризуется и закон постоянного относительного роста (см. формулу 2.1), только с той разницей, что в данном случае константа имеет отрицательное значение и закон выражает "постоянное относительное убывание" функции. Этот закон известен во многих областях науки. Оказывается, что закон Ципфа совпадает по форме с неким универсальным законом, охватывающим широкий круг явлений материального мира. В данном случае частота (F_i) убывает со скоростью, пропорциональной росту словаря, измеряемому особым образом (по рангам i).

Ранговое распределение, соответствующее функциям (2.8) и (2.9), имеет на графике форму гиперболы, а в билогарифмических координатах оно дает прямую линию, т.е. наблюдается линейная зависимость между $\ln F_i$ и $\ln i$. Это подтверждается,

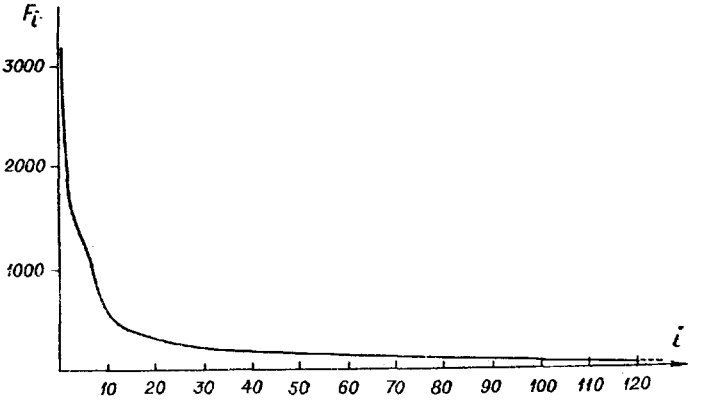


Рис. 2.1. Связь между частотой (F_i) и рангом (i) слова по данным ЧС словоформ эстонского языка.

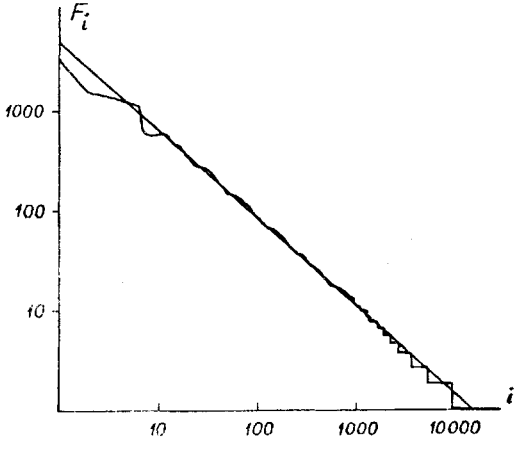


Рис. 2.2. Связь между F_i и i по данным ЧС словоформ эстонского языка. Билогарифмический масштаб.

например, на материале ЧС словоформ эстонского языка на основе текстов общим объемом около 100 тыс. словоупотреблений (см. рис. 2.2 по данным табл. 2.5).

Известно, что типологическое различие между языками обнаруживается особенно ярко при сопоставлении ранговых распределений словоформ. Сравнивая, например, данные ЧС словоформ эстонского и английского языков (при равных объемах текстов), мы можем констатировать, что значения параметра γ существенно различаются: для флективно-синтетического эстонского языка $\gamma = 0{,}86$, а для флективно-аналитического английского языка $\gamma = 1{,}002$ (Kučera H., Francis W.N., 1967, с. 357), т.е. угол наклона прямой на графике (и, соответственно, тангенс угла γ) для английского языка значительно больше, чем для эстонского языка (см. рис. 2.3). Это означает, что высокочастотные словоформы (в основном служебные слова) в английском языке покрывают бóльшую часть текста, чем в эстонском языке, и насыщение словаря словоформами осуществляется в английском тексте быстрее. В результате таких лингвистико-типологических расхождений в эстонском тексте объемом 100 тыс. словоупотреблений имеется около 30 тыс. словоформ, в то время как в английском тексте такого же объема их число не превышает 15 тыс.

Более подробный анализ показывает, что точная линейная зависимость между частотой и рангом (в билогарифмических координатах) выполняется не на всем протяжении рангового распределения. Во многих языках обнаруживается отклонение от линейной зависимости в ранговом распределении языковых единиц в области больших частот (в зоне "ядра" ЧС). Для улучшения соответствия между эмпирическими и теоретическими данными обычно используется вариант закона Ципфа, включающий т. наз. поправку Мандельброта (Mandelbrot B., 1954). Этот вариант называется "каноническим законом" Ципфа или "законом Ципфа-Мандельброта":

$$F_i = C(i+B)^{-\gamma} \quad \text{или} \quad p_i = k(i+B)^{-\gamma}, \qquad (2.11)$$

где B – поправка Мандельброта. В частном случае, если $\gamma = 1$, формула имеет вид

$$F_i = C(i+B)^{-1} \quad \text{или} \quad p_i = k(i+B)^{-1} \qquad (2.12)$$

Формула (2.11) дает хорошее соответствие между теоретическими и эмпирическими данными в начальной части ЧС, причем

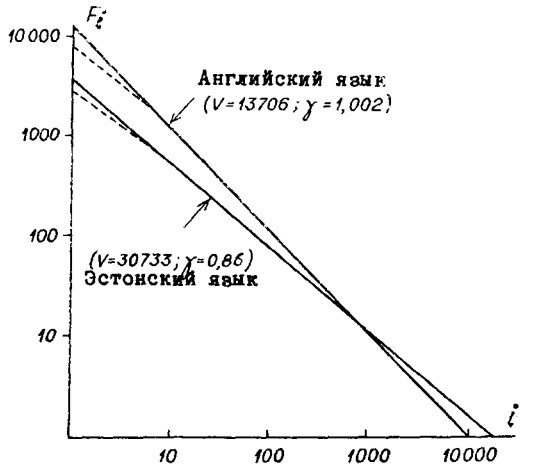


Рис. 2.3. Ранговое распределение: связь между частотой (F_i) и рангом (i) по данным ЧС словоформ эстонского и английского языков (выборки объемом N = = 100 000 словоупотр.; V - объем словаря). Билогарифмический масштаб.

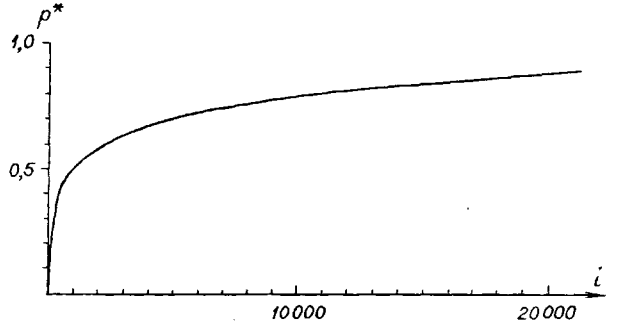


Рис. 2.4. Связь между накопленной относительной частотой (p_i^*) и рангом (i) по данным ЧС словоформ эстонского языка.

10

применение поправки Мандельброта приводит к некоторому увеличению теоретических значений констант C (или k) и γ (см. табл. 2.6).

Кроме отклонения от линейной зависимости (в билогарифмических координатах) в головной части ЧС, во многих случаях наблюдается более или менее заметное отклонение от линейной зависимости в области малых частот (в "хвосте" распределения). По данным сводного словаря эстонского языка такое отклонение небольшое (например, в ЧС лексем при $i = 30 \div 2500$ тангенс угла наклона $\gamma = 0{,}98$, а при $i > 2500$ $\gamma = 1{,}03$). Отклонение в области малых частот более заметное по данным сводных словарей русского языка (табл. 2.7), а так же по данным индивидуального словаря эстонского языка (табл. 2.8). Для иллюстрации приводится ранговое распределение частот лексем по данным ЧС русского языка (табл. 2.9) и соответствующий рисунок (рис. 2.5), на котором ясно видны отклонения в начальной и конечной частях распределения.

Отклонения от линейной зависимости (в билогарифмических координатах) в начальной и конечной частях рангового распределения, как правило, увеличиваются с нарастанием объема выборки (текста). Можно с полным правом утверждать, что частотная структура текста д и н а м и ч н а и закономерно изменяется с изменением объема текста. Такая динамика частотной структуры наблюдается в отношении как сводных, так и индивидуальных текстов.

Таким образом, отклонения от закона Ципфа в его основной форме можно считать закономерным явлением, учитывая динамический характер статистической организации текста. Представляется, что наиболее общим условием выполнения закона Ципфа в его ранговой форме (по крайней мере, для текстов не очень большого объема) следует считать линейность связи между $\ln F_i$ и $\ln i$ в средней части распределения, в то время как отклонения в начальной и конечной частях, а также конкретное значение параметра γ объясняются разными, в т.ч. лингвистическими причинами (тип языка, структурные особенности словаря, выбор единиц и др.; см. Борода М.Г., Поликарпов А.А., 1984). Кроме того, по "точкам перегиба" возможно разбиение словаря на объективно выделяющиеся частотные зоны.

Особой точки зрения придерживается Г.Я.Мартыненко (1978), который считает, что упомянутое распределение, как распределение неоднородных элементов, в принципе не может быть аппроксимировано единой функцией, в частности ядерные и пери-

Таблица 2.6

Значения параметров распределения Ципфа-Мандельброта k, γ, B по данным ЧС разных языков (наш материал; Бектаев К.Б., 1978; Калинина Е.А., 1968; Kučera H., Francis W.N., 1967)

Язык и подъязык	Тип единиц	Объем текста	Объем словаря	Параметры C	к	γ	B
Эстонский (худ. проза; сводный текст)	Лексемы	99898	14654	6793	0,068	0,92	-
				14785	0,148	1,04	2,3
	Словоформы	99898	30733	4095	0,041	0,86	-
				4595	0,046	0,87	0,5
Казахский (худ. проза; индив. текст)	Словоформы	105494	22642	6224	0,059	0,87	-
				8334	0,079	0,92	4,96
Русский (электроника)	Словоформы	100000	14062	4200	0,042	0,81	-
				5400	0,052	0,84	1,9
Английский (смешанные тексты)	Словоформы	101566	13706	13100	0,129	1,002	-
				14000	0,138	1,05	1,0

Таблица 2.7

Значения параметра γ в разных частотных зонах по данным ЧС эстонского и русского языков (I – Засорина Л.Н.,1966; II – ЧС русского языка, 1977)

Частотная зона (i)	Эстонский язык		Русский язык (лексемы)	
	лексемы	словоформы	I	II
I ÷ 30 (γ_1)	0,83	0,77	0,70	0,71
30 ÷ 2500 (γ_2)	0,98	0,89	0,94	0,95
> 2500 (γ_3)	1,03	0,90	1,24	1,51
Весь словарь (γ)	0,92	0,86	0,93	1,0
Объем словаря (L)	14654	30733	10830	39268
Объем текста (N)	99898	99898	120474	1056382

Таблица 2.8

Значения параметра γ в разных частотных зонах по данным ЧС лексем I-го тома романа "Правда и право" ("Tõde ja õigus") А.Х. Таммсааре (вычисления на основе данных: Villup A., 1978) – эст. яз.

Частотная зона (i)	Весь роман (I-й том)	В том числе	
		авторская речь	речь персонажей
I ÷ 30 (γ_1)	0,7	0,68	0,59
30 ÷ 1500 (γ_2)	1,1	1,11	1,21
> 1500 (γ_3)	1,4	1,43	1,47
Весь словарь (γ)	1,0	1,0	1,01
Объем словаря (L)	8228	7348	3135
Объем текста (N)	160356	114124	46232

Примечание: Значения параметров (табл. 2.7 и 2.8) вычислены на основе средних рангов методом наименьших квадратов. Средние ранги вычисляются при образовании "платформ", т.е. когда встречается ряд слов одинаковой частоты. Например, по данным ЧС русского языка (табл. 2.9) частоте $F = 1$ соответствуют ранги $i = 25890 \div 39268$; средний ранг $i \approx 32600$. (О преимуществе вычисления параметров на основе средних рангов см. Калинин В.М., 1964, с. 125).

Таблица 2.9

Ранговое распределение частот лексем по данным ЧС русского языка (1977); N = 1056382; L = 39268; F_{max} = 42854.

i	F_i	i	F_i	i (средн.)	F_i
1	42854	200	557	8750	10
2	36266	300	332	(8451-9045	
3	19228	400	312	9400	9
4	17261	500	256	(9046-9758)	
5	13839	600	217	10200	8
6	13307	700	187	(9759-10599)	
7	13185	800	164	11000	7
8	13143	900	147	(10600-11576)	
9	12975	1000	134	12200	6
10	10719	1500	90	(11577-12855)	
15	6246	2000	67	13700	5
20	5179	2500	52	(12856-14536)	
30	4156	3000	42	15650	4
40	2830	4000	30	(14537-16779)	
50	2136	5000	22	18500	3
60	1887	6000	17	(18780-20143)	
70	1655	7000	14	23000	2
80	1415	7300	13	(20144-25889)	
90	1210	7700	12	32600	1
100	1093	8200	11	(25890-39268)	

ферийные элементы описываются разными законами. Особый интегральный подход предлагает В. В. Нешитой (1984; 1986), который решает задачу аналитического описания различных сторон частотной структуры текста на основе оригинальной обобщающей системы непрерывных распределений.

Новые трактовки закона Ципфа. В связи с констатацией отклонений от линейной зависимости переменных (в билогарифмическом масштабе) представляет интерес тезис П. М. Алексеева (1978) о "нелинейном" характере рангового распределения лексических единиц в больших текстах. Предполагается, что при большом увеличении данного текста многие редкие слова передвигаются в средние ярусы частотного словаря и происходит изменение характера зависимости ранг - частота: платформы в нижних ярусах уменьшаются, зато они будут расширяться в средней части словаря, сдвигая эту часть теоретического графика вправо и увеличения его кривизну. П.М. Алексеев предлагает следующую обобщающую формулу закона Ципфа:

$$F_i = C i^{-(\beta + q \ln i)} \quad \text{или} \quad p_i = k i^{-(\gamma + q \ln i)}, \quad (2.13)$$

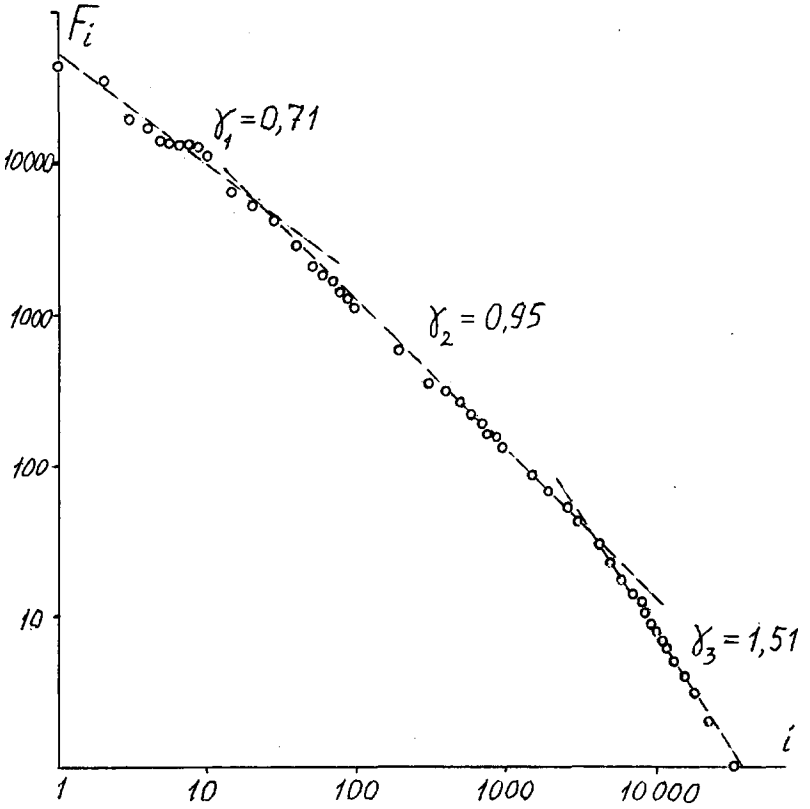


Рис. 2.5. Ранговое распределение: связь между частотой (F_i) и рангом (i) по данным ЧС лексем русского языка (1977). При малых частотах используются средние ранги (см. табл. 2.9). Билогарифмический масштаб.

где C (или k), γ и φ - параметры. Эта формула в логарифмической записи соответствует уравнению параболической регрессии. П.М. Алексеев называет выражение (2.13) "четвертым приближением" закона Ципфа, имея в виду, что первым приближением является классическая однопараметрическая форма (формула (2.9)), а вторым и третьим приближениями - формулы (2.8) и (2.11).

Формула (2.13) обладает тем преимуществом, что она включает в себя основные формы распределения Ципфа в качест-

ве частных случаев: при $\varphi = 0$ формула (2.13) превращается в формулу (2.8), а если при этом $\gamma = 1$, то получается однопараметрическая формула (2.9). Проверка показывает, что формула (2.13) хорошо аппроксимирует эмпирические данные в том случае, если имеются отклонения от "линейной" зависимости в начальной и конечной частях распределения. Вопрос только в том, отражает ли формула (2.13) действительную структуру текста (параболическую зависимость между F_i и i в билогарифмических координатах), или она лишь "затушевывает" факт закономерных отклонений в начальной и конечной частях распределения, игнорируя то, что в средней части распределения в действительности сохраняется линейность связи (на графике прямая линия в билогарифмических координатах). Для окончательного решения проблемы потребуются новые исследования на материале больших массивов текстов.

Дальнейшее теоретическое осмысление "нелинейной" концепции закона Ципфа дано в работе В. Н. Б ы ч к о в а (1984), который развивает тезис о "расщеплении" параметра на две составляющие - собственно константу γ_0 и переменную, "скользящую" γ_i. Формулу (2.13) можно переписать в виде

$$F_i = C i^{-\gamma_i}, \qquad (2.14)$$

где γ_i (т.е. значение γ-параметра в конкретной точке частотно-рангового ряда) определяется как

$$\gamma_i = \gamma_0 e^{di}, \qquad (2.15)$$

где d - коэффициент прироста γ-параметра за интервал перехода от $i = 1$ до $i_{max} = V$.

Основываясь на некоторых вероятностно-комбинаторных рассуждениях, Ю. К. К р ы л о в (1982) выводит формулу связи между частотой и рангом слова (в наших обозначениях)

$$F_i = \frac{C}{i + B_1} - B_2, \qquad (2.16)$$

где C, B_1 и B_2 - параметры. Легко видеть, что эта формула отличается от формулы Ципфа-Мандельброта при $\gamma = 1$ (2.12) лишь наличием слагаемого B_2. Параметры B_1 и B_2 обеспечивают возможность сдвига в направлениях, параллельных координатным осям, благодаря чему учитываются отклонения от "линейной" связи как в начальной, так и в конечной части распределения. Подразумевается, что ципфовский параметр γ, указывающий на уклон прямой в билогарифмических координатах,

равняется единице. В тех случаях, когда параметр γ по эмпирическим данным близок к единице, формула достаточно хорошо описывает связь между F_i и i.

Факт отклонения от линейной зависимости (в билогарифмических координатах) рангового распределения послужил поводом для выведения особого варианта закона Ципфа также польским исследователем Е. В о р о н ч а к о м (Woronczak J., 1967). В его формуле учитываются отклонения как в начальной, так и в конечной частях распределения, так же как и у Ю.К. Крылова, но в отличие от последнего Е. Ворончак предлагает вариант, в котором параметр γ (коэффициент убывания) может отличаться от единицы ($\gamma \neq 1$). В наших обозначениях формула Е. Ворончака имеет вид:

$$F_i = N(i+B)^{-\gamma} z^i \phi^{-1}, \qquad (2.17)$$

где γ, B и z - параметры, N - объем текста,

$$\phi = \sum_{i=0}^{\infty} (i+B)^{-\gamma} z^i \qquad (2.18)$$

Параметр B играет роль поправки Мандельброта, учитывающей отклонение в начальной части распределения, а параметр z (вернее, выражение z^i при $|z| < 1$) обеспечивает более быстрое убывание функции по сравнению с формулой Мандельброта (2.11); тем самым учитывается отклонение в конечной части распределения. По существу, множитель z^i имеет такое же действие, как и возрастание γ с рангом i. Этот принцип в некоторой степени роднит метод Е. Ворончака с "нелинейным" подходом П.М. Алексеева и В.Н. Бычкова, и подобно им Е. Ворончак по существу игнорирует "линейный" характер связи на среднем участке распределения.

Формула Мандельброта в своей двухпараметрической форме (2.12) является основой для "обобщенного закона Ципфа-Мандельброта" Ю. К. О р л о в а (1970; 1976; 1978). Этот автор развивает оригинальный подход, при котором центральным понятием является т.наз. "объем ципфа" (обозначаемый символом Z), служащий отправной точкой для вычисления параметров частотной структуры текста. Если исходить из теоретически вычисляемого значения Z (о вычислениях см. Орлов Ю.К., 1978), то формула принимает вид

$$p_i = k(i+B)^{-1}, \qquad (2.19)$$

где p_i - относительная частота слова с рангом i;

$k = [\ln(\bar{Z} p_{max})]^{-1}$; $B = k p_{max}^{-1} - 1$; p_{max} - относительная частота наиболее частотного слова.

Отметим, что параметр γ, указывающий на наклон прямой в билогарифмических координатах, в этой формуле жестко определен и равен единице ($\gamma = 1$). (О варианте с $\gamma \neq 1$ см. Орлов Ю.К., 1976, с. 184 и след.).

Формула Ю.К. Орлова дает достаточно хорошие результаты при условии, что параметр γ по эмпирическим данным близок к единице и вычисление "объема Ципфа" ($\bar{Z}$) для данного текста оказывается удачным. "Объем Ципфа" указывает по замыслу Ю.К. Орлова на тот единственный объем данного текста, при котором теоретически может выполняться закон Ципфа-Мандельброта, причем этот объем связывается с понятием "целостности" текста.

Все рассмотренные выше варианты формул закона Ципфа (в ранговой форме) представляют собой непрерывные функции, хотя в действительности мы имеем дело с дискретным распределением лингвистических объектов. Обычно игнорируется этот факт, так как считается, что непрерывная функция вполне адекватно отражает все особенности "скачкообразно" изменяющейся эмпирической функции частотно-рангового распределения. Желая все же уточнить вид распределения, М. В. Арапов и др. (1975) предлагают особый "дискретный аналог закона Ципфа", при котором учитывается дополнительное условие - целочисленность частоты или ранга. Соответствующая формула имеет вид:

$$F_i = \frac{\beta (L+1)^{\gamma}}{1-\gamma} \left[(i+1)^{1-\gamma} - i^{1-\gamma} \right], \quad (2.20)$$

где параметр β определяется в зависимости от значения ципфовского параметра γ (см. Арапов М.В. и др., 1975, с. 6), L - объем словаря (у авторов обозначение N).

В последнее время, гл. обр. благодаря исследованиям А.И. Яблонского (1976) и С.Д. Хайтуна (1983), утвердилось мнение, что закон Ципфа играет в соответствующих областях (в частности при описании социальных явлений) ту же универсальную роль предельного распределения, что и гауссов закон (нормальное распределение) в неорганических и др. процессах. В этой связи говорят о "негауссовости" закона Ципфа, содержательно выражающейся в явлении концентрации и рассеяния, а формально в отсутствии дисперсии (она равна бесконечности). Утверждается, что распределение Ципфа является не только одним из многих эмпирических распределений, а т е о р е т и -

ческим законом, имеющим надежную математическую базу в виде теории устойчивых "негауссовых" распределений (Яблонский А.И., 1976). Известен также тезис о том, что распределение Ципфа представляет собой универсальный закон, сферой действия которого являются "естественно возникшие сложные системы" (Аралов М.В., Шрейдер Ю.А., 1978, с. 75). К этому можно добавить, что, по мнению некоторых ученых, в распределениях типа "гиперболическая лестница" отражаются прежде всего системные свойства - целостность, организованность, многоуровневость (Козачков Л.С., 1978, с. 15).

Постоянство (устойчивость) распределения Ципфа применительно к социальным явлениям дает основание предположить, что явления, подчиняющиеся закону Ципфа, можно рассматривать как системы, находящиеся в равновесном, т.е. в наиболее благоприятном (оптимальном) для системы состоянии. Учитывая динамичность распределения и закономерную флуктуацию частот, ципфовскую структуру текста можно в целом охарактеризовать как подвижное равновесие системы.

Таким образом, и через системно-вероятностные рассуждения (негауссовость, подвижное равновесие) можно прийти к выводу об оптимизирующем характере закона Ципфа. В то же время, как нам представляется, надо остерегаться чрезмерного увлечения идеей оптимальности закона Ципфа, например, когда выполнение этого закона на языковом материале связывают с эстетическими свойствами текста, "литературной завершенностью", "художественной полноценностью" и т.п. Надо учесть то, что с помощью закона Ципфа "измеряется" только абстрактная, формально-частотная структура (упорядоченность элементов строения) текста, вне связи его с конкретным языковым заполнением. Таким образом, зависимость между частотной структурой и содержательным аспектом текста может быть только очень косвенной, опосредованной, над ее установлением лингвистам предстоит еще очень много поработать.

Особо заслуживают внимания попытки связать принцип ципфовского распределения потоков речи ("концентрация и рассеяние") с деятельностью мозга. Например, А.Н. Лебедев (1983) пытается объяснить квантитативные особенности порождения речи пространственно-временной организацией периодических процессов головного мозга. Исходя из предположения о кодировании образов слов "пакетами волн нейронной активности", он сначала по аналогии выводит формулу для определения "полного диапазона колебаний ранга слова" (q) и

затем приходит к выражению (в наших обозначениях):

$$F = CQ,$$

где $C = F_{max}$ и $Q = \frac{1}{q} \ln\left(1 + \frac{q}{i}\right)$; i – среднее значение ранга слова. Тем самым построена формула, параметры которой по замыслу ее автора имеют ясный психологический и физиологический смысл. Близость опытных и теоретических данных подтверждается в предварительном эксперименте (Лебедев А.Н., 1983, с. 16).

Сравнительный анализ разных вариантов закона Ципфа показывает, что в зависимости от особенностей текста можно с успехом использовать один или другой вариант для описания рангового распределения слов. Например, для ЧС словоформ эстонского языка (табл. 2.5) при малых отклонениях в головной и хвостовой частях распределения хорошо подходит основная формула (2.8) закона Ципфа, для ЧС лексем (табл. 2.4) с отклонением в головной части лучше подходит формула с поправкой Мандельброта (2.11) и т.д. (подробнее о результатах сравнительного анализа см. Тулдава Ю.А., 1985). Обнаруженные отклонения от эмпирических данных объясняются естественной флуктуацией значений параметров лингвистических распределений, которые при квантитативно-системном подходе следует рассматривать как в е р о я т н о с т н ы е с и с т е м ы с моментами устойчивости и вариативности. Но во всех случаях сохраняется общий вид распределения, характеризующийся концентрацией и рассеянием объектов.

Наконец, следует сказать еще о том, что ранговое распределение частот лексических единиц можно представить и в и н т е г р а л ь н о й (кумулятивной) форме, т.е. в виде зависимости между рангом (i) и накопленной частотой (F_i^* или p_i^*). На графике это дает кривую, показывающую возрастание с приближением к максимуму, т.е. к полному объему словаря (см. рис. 2.4). Такое интегральное распределение можно аналитически выразить с помощью соответствующих формул закона Ципфа. Формулы в интегральной форме, соответствующие формулам (2.8) и (2.9), имеют следующий вид (ср. Арапов М.В., 1981; Хайтун С.Д., 1983):

$$F_i^* = \frac{C}{\gamma - 1}\left[1 - (i+1)^{1-\gamma}\right] \quad \text{(при } \gamma \neq 1\text{);} \quad (2.21)$$

$$F_i^* = a - b \ln i \quad \text{(при } \gamma = 1\text{).} \quad (2.22)$$

Значения параметров C и γ равны значениям этих параметров, установленным для формул (2.8) и (2.9) соответственно. Параметры $a \approx C \ln(1+B)$ и $b \approx C/\log e$ (см. Хайтун С.Д., 1983, с. 71).

Если учесть и поправку Мандельброта (B), то интегральная форма распределения определяется следующим образом:

$$F_i^* = \frac{C}{\gamma - 1}\left[(1+B)^{1-\gamma} - (i+1+B)^{1-\gamma}\right] \quad \text{(при } \gamma \neq 1\text{)}; \quad (2.23)$$

$$F_i^* = C \ln \frac{i+1+B}{1+B} \quad \text{(при } \gamma = 1\text{)}. \quad (2.24)$$

Известна также возможность аппроксимации интегрального рангового распределения с помощью функции Вейбулла (Белоногов Г.Г., Новоселов А.П., 1971). Функция Вейбулла, применительно к ранговому распределению слов, имеет вид:

$$F_i^* = N\left(1 - e^{-ci^k}\right), \quad (2.25)$$

где N – объем текста, i – ранг слова, c и k – параметры, e – основание натуральных логарифмов.

Интегральное ранговое распределение позволяет определять п о к р ы в а е м о с т ь текста фиксированным массивом лексических единиц. Она зависит от типа лексических единиц. При изучении покрываемости текста л е к с е м а м и было обнаружено, что данные разных языков приближаются друг другу особенно в зоне среднечастотных слов. Можно согласиться с утверждением Р.М. Фрумкиной (1961) о том, что в большинстве языков (вне зависимости от типа языка) покрываемость текста 1500 наиболее частотными лексемами колеблется около 80 $\pm$ 10 %. Притом надо учесть динамику изменения покрываемости в зависимости от длины текста: чем длиннее текст, тем меньше (в среднем) относительная покрываемость в зонах средне- и малочастотных слов.

Данные о покрываемости текста лексемами выявляют некоторые важные моменты в речевой деятельности. Оказывается, например, что 10 первых по частоте лексем, составляющих незначительную долю словаря, покрывают в тексте объемом $N \approx 100\,000$ (табл. 2.4) почти 20 %; 1000 наиболее частотным лексемам (7 % словаря) соответствует около 70 % текста и т.д. Можно также показать (по данным табл. 2.4), что 25-процентное покрытие текста дают 22 лексемы, 50-процентное покрытие – 196 лексем и 75-процентное покрытие – 1213 лексем. Подобные

данные имеют значение для методики преподавания языков, а также для решения некоторых других прикладных задач.

Покрываемость текста словоформами считается одной из важнейших характеристик квантитативной типологии языков. Отмечается, что покрываемость текста словоформами тесно связана с морфологическим строем и построением слова в исследуемом языке (Бектаев К.Б., 1978). Исследования показали, например, что в индоевропейских языках для покрытия 50 % текста требуется в среднем от 80 до 200 словоформ, в то время как в агглютинативных тюркских языках для этого требуется 700 - 800 словоформ. К.Б. Бектаев представил следующую схему для типологической характеристики языков на основе покрываемости текста 100 наиболее частотными словоформами: агглютинативные языки 20 - 28 %, флективно-синтетические языки 24 - 42 %, флективно-аналитические языки 43 - 54 % и флективно-аналитические языки с элементами аморфности 48 - 60 %. Эстонский язык, имея соответствующую покрываемость 28,9 % (табл. 2.5), попадает в группу флективно-синтетических языков, находясь все же близко к агглютинативным языкам. В английском языке (при объеме текста $N \approx 100\,000$) покрываемость текста словоформами составляет 47,6 % (Kučera H., Francis W.N., 1967, с. 313). В русском языке при таком же объеме текста соответствующий процент - 32,5 (Калинина Е.А., 1968, с. 104-105).

Частотный спектр лексики. Упорядоченный ряд численностей слов с данной частотой образует "спектральное" распределение частот, или частотный спектр лексики (именуемый в квантитативной лингвистике также частотно-лексическим, или лексическим спектром). Частотный спектр может рассматриваться как на уровне словаря, так и на уровне текста, и в обоих случаях он может быть представлен в дифференциальной (некумулятивной) или в интегральной (кумулятивной) форме (пример частотного спектра с некоторыми сокращениями по данным законченного индивидуального текста см. табл. 2.10).

Частотный спектр отражает тот же основной принцип концентрации и рассеяния лексических единиц, который был обнаружен при рассмотрении рангового распределения лексики. Концентрация единиц появляется здесь в области редких частот: в словаре и тексте слова с частотой $F = 1$ образуют наиболее многочисленную группу, затем следуют группы слов с частотами $F = 2$, $F = 3$ и т.д. Однако частотный спектр имеет динамический характер. При увеличении текста уменьшается доля

одноразовых слов как в словаре, так и в соответствующем тексте. Например, по данным ЧС английского языка (Kučera H., Francis W.N., 1967) доля одноразовых слов составляет в выборках различной длины:

	N = 2000	N = 100 000	N = 1 000 000
в словаре	69,9 %	51,6 %	44,7 %
в тексте	28,3 %	7,0 %	2,2 %

Соответственно увеличивается общая доля более частых слов. Некоторые ученые полагают, что при увеличении текста наступает момент, когда частотный спектр, в частности количество одноразовых слов, стабилизируется (см. например, Williams C.B., 1970, с. 103). Однако опытные данные говорят о том, что при увеличении текста удельный вес редких слов неуклонно падает. Теоретически можно даже предположить, что в очень большом тексте ("в генеральной совокупности") одноразовые, двухразовые слова могут встретиться только в виде исключения (Пиотровский Р.Г., 1975, с. 108). Действие механизма порождения частотного спектра можно объяснить следующим образом: при последовательном увеличении текста все большее число "нуль-частотных" слов, т.е. слов, реально бытующих в лексике исследуемого языка, но не попавших в выборку, попадает в выборочную совокупность и фиксируется в ЧС. Одновременно увеличивается частота некоторых редких слов, которые передвигаются в средние зоны ЧС. На характер изменения состава частотного спектра может повлиять также переход слов из средней зоны в зону высокочастотных слов и наоборот, хотя такие передвижения оказываются менее обычными при увеличении объема однородного текста (Пиотровский Р.Г., там же). В общем наблюдается закономерность "неравномерного перехода" (Muller Ch., 1975, с. 144): всякое подмножество m_i, т.е. число слов с частотой (i = 1, 2, ...) имеет тенденцию при увеличении текста "выигрывать" больше, чем "терять". Это означает, например, что переходы из группы m_i в m_{i+1} более многочисленны, чем переходы из группы m_{i+1} в m_{i+2}.

Что касается аналитического выражения спектрального распределения частот слов, т.е. частотного спектра лексики, то надо иметь в виду, что это распределение "органически" связано с ранговым распределением (они образуют две взаимосвязанные половины общей частотной структуры текста). Следовательно, сфера действия з а к о н а Ц и п ф а должна распространяться и на спектральное распределение. Если исходить из того, что ранговая форма данного распределения частот описы-

Таблица 2.10

Частотный спектр лексики по данным ЧС лексем романа А.Х. Таммсааре (зст.): F - частота слова; m - численность слов с данной частотой; p - относит. численность; m^*, mF^*, p^* - накопленные численности

F	Словарь				Текст			
	m	m^*	p	p^*	mF	mF^*	p	p^*
I	3637	3637	0,442	0,442	3637	3637	0,023	0,023
2	I2I6	4853	0,148	0,590	2432	6059	0,0I5	0,038
3	6I3	5466	0,074	0,664	I839	7908	0,0II	0,049
4	44I	5907	0,054	0,718	I764	9672	0,0II	0,060
5	297	6204	0,036	0,754	1485	II157	0,0I0	0,070
6	2I9	6423	0,027	0,78I	I3I4	I2471	0,008	0,078
7	I75	6598	0,02I	0,802	1225	13696	0,007	0,085
8	124	6722	0,0I5	0,8I7	992	14688	0,006	0,092
9	1I0	6832	0,0I3	0,830	990	I5678	0,006	0,098
I0	85	6917	0,0II	0,84I	850	I6528	0,005	0,I03
II-20	446	7363	0,054	0,895	6609	23I37	0,04I	0,I44
2I-50	42I	7784	0,05I	0,946	I3826	36963	0,086	0,230
5I-I00	193	7977	0,024	0,970	I3I53	50II6	0,082	0,3I2
I0I-500	202	8I79	0,024	0,994	42576	92692	0,266	0,578
50I-1000	29	8208	0,0035	0,9975	20841	II3533	0,I30	0,708
> I000	20	8228	0,0025	I,0	46823	I60356	0,292	I,0
Σ	8228 (L)	-	I,0	-	I60356 (N)	-	I,0	-

вается формулой закона Ципфа с поправкой Мандельброта (см. формулу (2.II)), то спектральным аналогом рангового распределения является (см. Хайтун С.Д., 1983, с. 161):

$$m(F) = c\,F^{-(1+\alpha)}, \qquad (2.26)$$

где $m(F)$ - число слов с частотой F; c и α - параметры, причем $\alpha = 1/\gamma$ (γ из формулы (2.II)) и $c = \alpha(L-1)/(1-F_{max}^{-\alpha})$, где L - объем словаря, F_{max} - частота наиболее частотного слова. Так как в распределениях частот слов $F_{max}^{-\alpha}$ обычно минимально мало и $L \gg 1$, то практически можно считать, что $c \approx \alpha L$.

По теоретическим соображениям можно предположить, что соответствие между параметром γ рангового распределения и параметром α из формулы (2.26) достигается наилучшим образом в том случае, если исходить из значений параметра γ, установленных для средней или хвостовой части рангового распределения. Эти части рангового распределения соответствуют средней и начальной частям спектрального распределения частот слов. Как известно, параметр γ (выражающий уклон линии рангового распределения в билогарифмическом масштабе) обычно меняет свое значение в соответствии с отклонениями в начальной и хвостовой частях рангового распределения. В общих чертах можно говорить о трех "стадиях" распределения, которым соответствуют параметры γ_1, γ_2, γ_3 (для начальной, средней и хвостовой части).

При $\gamma = I$ формула принимает вид

$$m(F) = c\,F^{-2}. \qquad (2.27)$$

Такой представил себе эту зависимость первоначально сам Ципф, хотя он позднее предложил уточненный вариант:

$$m(F) = c\,(F^2 - 0{,}25)^{-1}. \qquad (2.28)$$

Нами была предложена формула с добавочным параметром (d) типа поправки Мандельброта (Тулдава Ю.А.,1986а, с.148):

$$m(F) = c\,(F+d)^{-\beta}, \qquad (2.29)$$

где c, d и β - параметры. Формула совпадает по форме с законом Ципфа-Мандельброта для ранговых распределений.

Были еще другие попытки более точно аппроксимировать спектральное распределение прибавлением добавочных параметров, например, по следующей формуле (Krallmann D., 1966,

с. 88):

$$m(F) = c F^{-k} e^{bF}, \qquad (2.30)$$

где c, b и k - параметры.

Основываясь на том или другом варианте закона Ципфа, еще другие авторы вывели свои формулы для аналитического выражения частотного спектра (Орлов Ю.К., 1976; Арапов М.В. и др., 1975; Крылов Ю.К., 1982; Brookes B.C., 1982; и др.). Большой интерес представляет новейшая попытка Ю.К. Крылова (1987) теоретически вывести формулу частотного спектра, исходя из некоторых вариационных принципов. Его формула в принципе сходна с формулой (2.30), но имеет другое обоснование:

$$m(F) = c F^{-2} e^{b/F}, \qquad (2.31)$$

где c и b - параметры, которые вычисляются теоретически на основе наблюдаемых величин L (объем словаря), N (объем текста) и F_{max} (частота наиболее частотного слова в данном тексте).

Все упомянутые формулы с большей или меньшей точностью описывают спектральное распределение частот слов в словаре и тексте. Проиллюстрируем применение двух формул (2.26) и (2.29) на материале лексем разных языков (табл. 2.11). Оказывается, что при отклонении от "линейности" в начальной части распределений лучшие результаты дает формула (2.29) с поправочным коэффициентом.

Основная формула (2.26), соответствующая закону Ципфа в ранговой форме, представляет собой такую же степенную функцию с отрицательным показателем степени, как и функция для выражения рангового распределения. В дифференциальной записи, раскрывающей сущность изменения переменных в отношении друг друга, обнаруживается, что как ранговая, так и спектральная формы распределения Ципфа подчиняются закону "постоянного относительного роста или убывания", при котором наблюдается постоянное отношение между относительными приростами функции и аргумента (пропорциональность относительных приростов). Для частотного спектра имеем:

$$\frac{dm/m}{dF/F} = -(1+\alpha). \qquad (2.32)$$

Частичные отклонения от такой простой и естественной зависимости объясняются разными, в т.ч. лингвистическими причинами.

Таблица 2.II

Частотные спектры: наблюдаемое и ожидаемое количество слов $m(F)$ с частотой F по данным 1) ЧС лексем романа А.Х. Таммсаре (эстонский язык); N = 160 356, L = 8228; 2) ЧС лексем русского языка (1977); N = 1 056 382; L = 39 268. Вычисления по формулам: I (2.26): $m(F) = cF^{-(1+\alpha)}$; II (2.29): $m(F) = c(F+d)^{-\beta}$ III (2.33): $m_{i+1} = m_i (a+i-1)/(x+i)$

F	ЧС лексем романа А.Х. Таммсаре (вст.)				ЧС лексем русского языка (1977)			
	$m(F)$ набл.	$m(F)$ о ж и д. I	II	III	$m(F)$ набл.	$m(F)$ о ж и д. I	II	III
1	3637	5700	3617	(3637)	13379	17000	13068	(13379)
2	1216	1754	1295	1436	5746	6010	5771	6690
3	613	881	676	763	3364	3272	3368	3983
4	441	540	420	472	2243	2125	2253	2638
5	297	370	288	320	1681	1521	1634	1872
6	219	271	211	231	1279	1157	1251	1395
7	175	208	162	174	977	918	995	1078
8	124	166	129	136	841	751	815	867
9	110	136	105	109	713	630	682	698
10	86	114	87	89	595	538	580	578
15	53	57	43	41	286	293	311	275
20	34	35	26	23	200	190	199	159
30	15	18	13	10	109	104	105	70
40	7	10	7	7	60	67	67	40
50	4	7	5	5	45	48	47	30
60					30	37	35	20
70					30	29	28	
80	3	3	2	3	26	24	22	
90					15	20	19	
100	2	2	2	2	14	17	16	
150					7	9	8	
200	1	1	1	1	4	6	5	
300					3	3	3	
400					2	2	2	
500					2	2	1	
600					1	1	1	
Параметры:		= 5700	= 5800	= 1,35	-	= 17000	= 25000	= 2,08
		= 0,7	= 1,8	= 2,42		= 0,5	= 1,6	= 3,16
			= 0,3				= 0,5	

Примечание: Для частот $F \geq 100$ взяты средние значения $m(F)$; например, для F = 100 вычислено среднее значение $m(F)$ в промежутке F = 98 ÷ 102.

Модель частотного спектра У э р и н г а - Х е р д а - н а (см. Herdan G., 1964) представляет собой другую возможность аналитического описания спектрального распределения частот слов. Здесь исходят из представления о том, что спектральное распределение частот слов образует монотонно убывающий ряд $m_1, m_2, \ldots m_n$ (т.е. число слов с частотой 1, частотой 2 и т.д.), который определяется двумя параметрами a и x:

$$m_{i+1} = m_i \frac{a+i-1}{x+i} \qquad (2.33)$$

где i - частота слова (i = 1, 2, ...). Эта модель соответствует закономерности "неравномерного перехода", о которой шла речь выше. Практически это означает постоянное уменьшение отношения m_i/m_{i+1}. Для применения модели нужно предварительно знать объем текста (N), объем словаря (L или V) и число одноразовых слов (m_1). Параметры a и x вычисляются следующим образом: $a = (Q-M-1)^{-1}$; $x = aQ$ где $Q = (1 - m_1/L)^{-1}$; $M = L/N$.

Считается, что модель Уэринга-Хердана работает хорошо на выборках умеренного объема ($N < 200\,000$). Это подтверждается на материале индивидуального текста ($N \approx 160\,000$), в то время как в большом сводном тексте (при $N \approx 1$ милл. СУ) соответствие хорошее только в начальной части частотного спектра, примерно до m_{15}, но в дальнейшем модель дает сильно сниженные оценки спектра (см. табл. 2.11).

Модель Уэринга-Хердана замечательна тем, что она показывает переходы от m_1 к m_2, от m_2 к m_3 и т.д. На этом основании французская исследовательница Дольфэн (Dolphin, 1974; цит. по: Muller Ch., 1976) высказала гипотезу, что по этой модели можно будет определить и переход назад, в частности от m_1 до m_0, где под m_0 подразумеваются "нуль-частотные" слова, т.е. слова, которые (по предположению) относятся к лексикону автора или авторов, но которые в данном тексте не используются. Дольфэн предлагает на этот случай особый метод вычисления параметров a и x (Muller Ch., 1976, с. 143). Но и при традиционном способе можно вычислить m_0 следующим образом: по формуле (2.33) получается $m_1 = m_0 \frac{a-1}{x}$ и, следовательно, $m_0 = m_1 : \frac{a-1}{x}$. Предварительные опыты показывают, что этот метод можно успешно использовать при стилеметрическом анализе выборок приблизительно одинакового объема, причем показатель m_0 истолковывается как особая стиледифференцирующая характеристика.

Спектральное распределение частот слов может быть представлено также в интегральной (кумулятивной) форме. На основе данных о ранговом распределении по формуле Ципфа (2.8) конструируется следующая формула интегрального спектрального распределения (Хайтун С.Д., 1983, с. 161):

$$m^*(F) = L\,(1 - F^{-\alpha}), \qquad (2.34)$$

где $m^*(F)$ - накопленные численности слов, F - частота слова, L - объем словаря, параметр $\alpha = 1/\gamma$. Здесь так же, как и в дифференциальном распределении, требуется ввести поправочный параметр (c) для улучшения соответствия между эмпирическими и теоретическими данными:

$$m^*(F) = L\left[1 - (F + c)^{-\alpha}\right]. \qquad (2.35)$$

Интегральный частотный спектр дает представление о покрываемости текста словами разных частот. Такие данные имеют значение при стилометрическом анализе текстов и при типологическом сравнении языков.

Особый интерес представляет то обстоятельство, что интегральный частотный спектр слов хорошо описывается логнормальным распределением (см., например, Carroll J.B., 1967), если довольствоваться т.н. усеченной формой логнормального распределения (подробнее см. Тулдава Ю.А.,1986; Манасян Н.С., 1987). Как известно, логнормальное распределение используется для описания определенного рода вероятностных процессов. В этом смысле распределение частот слов (частотный спектр) может быть интерпретировано как результат какого-то особого вероятностного процесса, действующего при порождении речи. Такого мнения придерживаются многие исследователи, причем считается, что логнормальность распределения отражает "присущий естественному языку принцип оптимального кодирования информации" (Herdan G., 1964, с. 61-62). Однако имеются и возражения против использования логнормального распределения для аппроксимации спектрального распределения частот слов. Считается, что аппроксимация некорректна в силу того, что в данного типа лингвистических распределениях наблюдается зависимость моментов от объема выборки, а это противоречит натуре "гауссовых" распределений, к которым относится логнормальное распределение (см. Хайтун С.Д.,с. 81 и 184-185).

В итоге можно констатировать, что многие из рассмотрен-

ных методов позволяют вполне удовлетворительно аналитически описывать спектральное распределение частот слов. Такая множественность решений не должна никого удивлять. В науке уже давно известно, что математическими моделями изучаемых явлений могут служить различные функции, задаваемыми уравнениями экспоненты, гиперболы, логисты и т.д., а все дело в правомерности принятой системы исходных постулатов по отношению к конкретному явлению. Предпочтение одной или другой модели зависит от содержательного анализа конкретной проблемы и возможности адекватной интерпретации модели. В нашем конкретном случае можно, например, предпочесть модель Ципфа по той причине, что она выражает очень простую и естественную зависимость между переменными (по закону "постоянного относительного роста или убывания" переменных) или что она в некотором смысле указывает на связь с деятельностью мозга (гипотеза А.Н. Лебедева). В качестве альтернативы (или дополнения) можно принять модель Уэринга-Хердана по той причине, что здесь явно просматривается связь с закономерностью "неравномерного перехода", наблюдаемого на практике, причем есть возможность учитывать наличие т.н. нуль-частотных слов. Принятие модели логнормального распределения позволяет рассматривать порождение речи в качестве вероятностного процесса, что может служить основанием для определенных выводов о природе языка и т.п.

2.3. ЗАВИСИМОСТЬ "СЛОВАРЬ - ТЕКСТ"

Постановка вопроса. Вопрос о количественной зависимости между объемом словаря и объемом текста в динамике порождения речи имеет как теоретическое, так и практическое значение в квантитативной лингвистике. Если удастся смоделировать процесс нарастания объема словаря в зависимости от увеличения объема текста в виде некой ф у н к ц и и, имеющей определенный содержательный смысл, то это может не только расширить наши знания о наиболее общих квантитативных закономерностях порождения речи, но и позволит решить ряд интересных и актуальных задач прикладного характера. На основе функции, выражающей зависимость объема словаря $(L)^+$ от объема текста (N), можно, например, находить

+ В зависимости от конкретных условий исследования объем словаря может быть выражен количеством лексем (L) или словоформ (V).

неизвестное значение L по данному N, или наоборот, а также определить степень насыщения или достаточность объема выборки при проектировании автоматизированных информационных систем. Установление формы связи между объемом словаря и объемом текста позволяет также исследовать стилистические особенности индивидуальных текстов или жанров и содействует решению некоторых педагогических и психологических задач (измерение сложности текста, определение лексического "богатства" текста, установление авторства и др.).

Имеются многочисленные попытки построения эмпирических формул для выражения зависимости объема словаря от объема текста (например, Kuraszkiewicz W., 1958; Guiraud P., 1959; Somers H., 1959; Müller W., 1971; Захарова А., 1967). Наряду с применением чисто эмпирических формул были попытки смоделировать процесс нарастания словаря, исходя из определенных теоретических предпосылок, например, основываясь на предположении о логнормальном распределении слов в тексте (Carroll J.B., 1967) или о действии закона Ципфа (Калинин В.М., 1964; Херц М.М., 1969). Основой для выведения формул роста словаря использовались и другие известные распределения, например, распределение Вейбулла (Нешитой В.В., 1975; 1984). Основываясь на нейрофизиологическом механизме процедуры выбора, А.Н. Лебедев (1986) развил теорию о связи порождения речи с особенностями памяти человека и предложил модель определения объема словаря по объему текста, количественно, с привлечением единых нейрофизиологических параметров. На основе анализа определенных свойств интегральных функций распределения Ю.К. Крылов (1985) построил особую модель зависимости "словарь - текст", связанную с изменением численности одноразовых слов и ранговым распределением частот слов в тексте.

Считается, что теоретически наиболее приемлемыми являются модели, которые позволяют рассматривать зависимость между объемом словаря и объемом текста в связи с другими сторонами статистической организации текста. Поэтому, важным событием в квантитативной лингвистике было появление работ В.М. Калинина (1964 и 1965), в которых была впервые решена задача построения комплексной модели, включающей собственно частотную структуру текста (ранговое и спектральное распределения частот) и зависимость "текст - словарь". Однако ввиду сильной идеализации структуры реального текста (допущение о случайности и независимости появления слов в тексте и

жесткий характер взаимосвязи разных квантитативных аспектов текста) эту модель редко удавалось применять на практике при лингвистическом анализе текста. Несколько лучших результатов добился Ю.К. Орлов (1978а), который предложил усовершенствованный вариант модели на основе т.н. "обобщенного закона Ципфа-Мандельброта" (критический анализ этого подхода при исследовании зависимости "текст - словарь" см. Тулдава Ю.А., 1980, с. 116-117).

Несмотря на теоретическую значимость "комплексных" (или "структурных") моделей, практическое их использование не всегда дает надежные результаты, особенно в тех случаях, когда требуется бóльшая точность, в частности при решении задач стилеметрического сравнения, прогноза и т.п. Известно, что в математических моделях обычно стремятся отразить как можно больше процессов в их связях и взаимозависимостях. Но один из видных специалистов по методологии математики пишет: "Однако еще не удалось построить модель, гармонично отражающую совокупность совместно протекающих процессов. Почти всегда выделяется для подробного отражения лишь один из процессов, об остальных используют лишь упрощенные данные" (Рыбников К.А., 1979, с. 110). Поэтому, можно попытаться исследовать зависимость "текст - словарь" с точки зрения "имманентных" свойств этого явления с целью построения более точной модели для ее использования при решении практических задач прогноза или стилеметрического анализа.

Модель последовательного выбора. Связь словаря и текста предстает в такой модели как исходный, первичный аспект статистической организации текста. Квантитативную зависимость между словарем и текстом следует при таком подходе изучать в связи с некоторыми наиболее общими факторами порождения текста. Мы основываемся при этом на предпосылках теории вероятностных систем (см. гл. I.I), согласно которым исследуемый объект характеризуется не только случайностью параметров (низший уровень организации), но и определенной устойчивостью и регулярностью в массе случайных событий (высший уровень организации).

В данном конкретном случае нарастание объема словаря в ходе порождения текста можно представить себе как вероятностный процесс, при котором на каждом шагу по ходу порождения текста осуществляется в ы б о р между "новым", ранее не появившимся словом, и "старым", уже употребленным в данном тексте словом. Эмпирически установлено, что с увеличением

текста вероятность выбора "нового" слова постоянно уменьшается. Отсюда можно заключить, что рассматриваемый вероятностный процесс подчиняется какой-то определенной глубинной закономерности порождения текста. Содержательно процесс нарастания словаря определяется сложным взаимодействием двух противопоставленных тенденций: стремлением к расширению и стремлением к ограничению словарного состава данного текста. С одной стороны, в процессе говорения (писания) говорящим (пишущим) овладевает желание развертывать и расширять выбранную тему (в силу ассоциативных свойств мышления). С другой стороны, свободное развертывание речи ограничивается свойствами человеческой памяти, а также, например, необходимостью оставаться в рамках определенной тематики, что ведет к повторению уже использованных знаменательных слов; к этому прибавляется постоянное повторение структурных (служебных) слов, обусловленное системно-языковыми причинами. Эти две тенденции определяют вероятностный процесс нарастания словаря в целом, однако в рамках нашего анализа следует более четко сформулировать общий структурно-функциональный принцип, который сохранял бы ясный лингвистический смысл и в то же время мог бы быть подвергнут математическому анализу. Регулирующим рычагом и движущей силой при порождении речи можно считать принцип о г р а н и ч е н и я р а з н о о б р а з и я л е к с и к и в тексте, имеющий, по-видимому, своей более глубокой причиной некоторые филогенетические факторы (аналогом этого принципа можно рассматривать принцип уменьшения энтропии в открытых самоорганизующихся системах).

Разнообразие лексики в тексте - несколько абстрактное понятие, но оно становится четким и наглядным, если его выразить с помощью известных в квантитативной лингвистике мер - "коэффициента разнообразия", т.е. отношения объема словаря к объему текста (L/N), или обратного отношения (N/L), которое выражает среднюю частоту слова в данном тексте. Известно, что степень разнообразия лексики изменяется вместе с увеличением объема текста: коэффициент разнообразия L/N (именуемый также "индексом TTR" от англ. type-token ratio) монотонно уменьшается, а средняя частота (повторяемость) слова соответственно увеличивается. Приводим для примера данные об изменении степени разнообразия лексики на материале "Капитанской дочки" А.С. Пушкина (см. Фрумкина Р.М., 1960):

N	L	L/N	N/L
5000	1568	0,31	3,19
10000	2432	0,24	4,11
29345	4900	0,17	5,99

Степень разнообразия лексики, измеряемая отношением между L и N, удобна для выявления некоторых существенных квантитативных свойств текста и в том смысле, что она (степень разнообразия) тесно коррелирует с вероятностным процессом выбора "нового" или "старого" слова на каждом шагу порождения текста. Отношение L/N отражает в определенном смысле вероятность появления (прибавления к словарю) нового слова. Действительно, если на достигнутом объеме текста в N словоупотреблений накоплен словарь из L различных слов, то вероятность прибавления нового слова к словарю пропорциональна отношению L/N.

Для выяснения некоторых дополнительных структурных свойств изучаемого явления можно представить процесс нарастания словаря как постоянное изменение "доли" использованного словаря в каком-то пространстве максимальных возможностей. Если, например, изобразить ход роста словаря графически (рис. 2.6), то можно наблюдать постоянно замедляющееся нарастание объема словаря по сравнению с "базисной прямой", которая указывает на тот идеальный объем словаря, при котором слова в тексте совсем не повторялись бы ($L = N$). Такой случай встречается в действительности в самом начале порож-

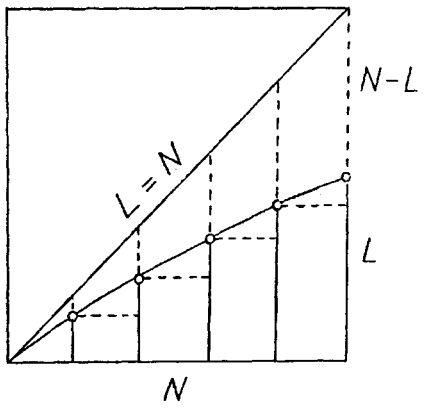


Рис. 2.6. Нарастание объема словаря (L) в зависимости от увеличения объема текста (N).

дения текста, а последующие этапы отражают неуклонный уход от начального состояния, где степень разнообразия имеет свое максимальное значение ($L/N = N/L = 1$). Как уже отмечалось, фактором, регулирующим этот уход, является обусловленный языковыми причинами процесс ограничения разнообразия лексики в реальном тексте.

В этой связи введем понятие д а в л е н и я р е к у р р е н т н о с т и (давления повторяемости слов), которое вместе с увеличением текста постепенно усиливается (отношение $(N-L)/N$, см. рис. 2.6), ограничивая степень разнообразия (отношение L/N). Если обозначить $L/N = p$, то давление рекуррентности (аналог избыточности) выражается через $q = (N-L)/N = 1 - L/N$; $p+q = 1$. Если учесть, что объем словаря L можно представить как некоторую долю "максимального" словаря, то исходя из вышесказанного, вычисление объема словаря можно осуществить на основе общих моделей:

$$L = Np \quad \text{или} \quad L = N(1-q).$$

Конструирование и проверка формул. Для наглядного представления возможностей различных решений проблемы аналитического выражения связи между объемом словаря и объемом текста применим индуктивный метод последовательной проверки гипотез и постепенного приближения к искомой форме зависимости на основе общей модели $L = Np$. Преимущество такого подхода в том, что мы не теряем связь с нашими исходными посылками о роли фактора разнообразия в процессе порождения текста и можем конструировать соответствующие формулы на единой системной основе.

В первом приближении можно исходить из предположения, что изменение с т е п е н и р а з н о о б р а з и я лексики – это непрерывный процесс, при котором движение происходит равномерно и прямолинейно. Эмпирическая проверка показывает, однако, что зависимость между $p = L/N$ и N можно аппроксимировать л и н е й н о й связью лишь на отдельных отрезках текста. Тем не менее можно на этой основе конструировать формулы, которые могут пригодиться при исследовании малых выборок, например, (Тулдава Ю.А., 1974):

$$L = Na(N+b)^{-1}, \qquad (2.36)$$

где a и b – параметры (причем a указывает на предел объема словаря). В данном случае подразумевается линейная связь между N/L (средней частотой слова) и N (объемом текс-

та).+

Теоретически более обоснованным является предположение о степенной связи между $\rho = L/N$ и N. В таком случае наблюдается линейная связь между $\ln(L/N)$ и $\ln N$, и зависимость между степенью разнообразия и объемом текста выражается функцией типа

$$L/N = aN^{b}, \qquad (2.37)$$

где a и b – параметры (в данном случае, когда L/N уменьшается при увеличении N, параметр $b < 0$). Эта формула по своей форме аналогична закону Ципфа (степенная функция с отрицательным показателем степени). Из этой зависимости получается

$$L = N(aN^{b}) = aN^{B}, \qquad (2.38)$$

где $B = b + 1$. В дифференциальной записи этой функции соответствует $(dL/L)/(dN/N) = B$, т.е. известный закон "аллометрического" или "постоянного относительного роста" (ср. формула (2.1)).

Как известно, Г. Хердан (Herdan G., 1966) считал функцию (2.38) универсальной для выражения связи между объемом словаря и объемом текста. Проверка показывает, что эта функция хорошо описывает связь между L и N в начальной стадии порождения текста (с начала текста до 5 – 10 тыс. словоупотреблений, т.е. в рамках одного короткого рассказа). В дальнейшем темп нарастания объема словаря в реальных текстах замедляется, и прогнозы по аллометрической функции дают завышенные оценки (подробнее см. Тулдава Ю.А., 1980).

Учитывая тот факт, что с увеличением текста темп нарастания объема словаря постепенно замедляется по сравнению с аллометрическим законом роста, можно прологарифмировать переменные. Анализ показывает, что целесообразно возвратиться к исходному принципу и определить закономерность нарастания объема словаря через "степень разнообразия лексики", т.е. с помощью логарифмирования переменных в формуле (2.37). В результате получаем

$$\ln(L/N) = a(\ln N)^{b}. \qquad (2.39)$$

+ Формулу (2.36) можно переписать в виде $N/L = \frac{N+b}{a}$, отсюда $N/L = \frac{1}{a}N + \frac{b}{a}$. Приняв $\frac{1}{a} = \alpha$ и $\frac{b}{a} = \beta$, получим линейное уравнение $N/L = \alpha N + \beta$.

Это дает нам $L/N = e^{a(\ln N)^b}$ $(a<0)$ и формулу роста словаря:

$$L = Ne^{-a(\ln N)^b} \qquad (2.40)$$

Важно отметить то обстоятельство, что функция (2.40) удовлетворяет "правильным" граничным условиям в начале текста, независимо от конкретных значений параметров a и b, т.е. при $N = 1$ объем словаря $L = 1$:

$$L(N=1) = 1e^{-a(\ln 1)^b} = 1e^0 = 1.$$

При практическом использовании формулы можно вычислить значения параметров a и b на основе линеаризации:

$$\ln\left|\ln\frac{L}{N}\right| = A + b\ln\ln N, \qquad (2.41)$$

где $A = \ln|a|$. Линейная связь (2.41) оказывается хорошей при анализе малых и больших текстов как индивидуальных, так и сводных (см., например, рис. 2.7; ср. табл. 2.12).

О возможностях экстраполяции. Хорошая описательная сила какой-нибудь функции в пределах опытных данных и при многократном прогнозе как вперед, так и назад наводит на мысль, что функцию можно использовать для экстраполяции на более дальние участки текста. Безусловно, это несколько рискованное предприятие, особенно если учесть, что мы в настоящее время не располагаем возможностями контроля таких прогнозов.

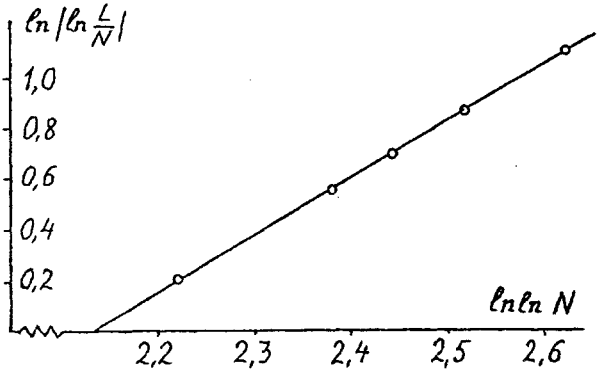


Рис. 2.7. Линейная связь между $\ln|\ln L/N|$ и $\ln\ln N$ (по данным ЧС английского языка Х.Кучера и В.Френсиса)

Но тем не менее практика автоматической обработки текстов может предъявлять требование хотя бы приблизительно прогнозировать объем словарей крупных текстовых массивов. В таком случае правомерно использовать именно такую формулу, которая показала себя на практике наиболее стабильной и достоверной. Такой формулой можно на основании наших опытов считать формулу (2.40), хотя формула имеет верхний предел (для текстов естественного языка при $N = 10^9 \div 10^{10}$).+

Экстраполяция, или п р о г н о з, основывается на предположении о "сохранении в основных чертах взаимосвязей прогнозируемого явления с другими явлениями" (Четыркин Е.М., 1975, с. 57) и может быть проведена только в условиях однородности явления. Под однородностью следует в данном случае понимать качественную однородность, т.е. одинаковость состава, например, одинаковую тематику текстов и соответственный словарный запас. При анализе сводных (смешанных) текстов требование однородности означает сохранение состава определенных исходных пропорций отдельных жанров или подъязыков и соблюдение прочих условий эксперимента.

Для примера приводим данные о динамике роста словаря лексем в творчестве А.С. Пушкина (Фрумкина Р.М., 1960):

	N	L		N	L
(1)	5000	1568	(3)	29345	4900
(2)	10000	2432	(4)	544777	21197

На основе данных по трем первым точкам (по данным романа "Капитанская дочка") вычисляем методом наименьших квадратов значения параметров a и b по формуле (2.40). В результате получаем $a = 0{,}0087$ и $b = 2{,}29$. Экстраполируя по формуле на точку $N = 544777$ (все творчество писателя), получаем $L = 22028$ т.е. близкий к реальным фактам результат.

Проверим эффективность формулы (2.40) также на материале сводного частотного словаря английского языка (Kučera H., Francis W.N., 1967). Вычисляем значения параметров на основе двух точек (при $N = 101566$ и $N = 253538$). Прогноз вперед и назад оказывается хорошим (см. табл. 2.12).

+ Анализ показал, что формула (2.40) в принципе подходит и для выражения зависимости между количеством одноразовых слов и объемом текста (см. Tuldava J., 1977).

Таблица 2.I2

Прогноз роста словаря по формуле (2.40) по данным ЧС английского языка (объем словаря в словоформах - V). Параметры a = 0,00891; b = 2,215.

	N	V	
Экспериментальные точки:	I0I566	I3706	
	253538	23655	
Прогноз вперед:	I0I4232	50617	(набл. 50406)
	I0000000	I48000	(-)
Прогноз назад:	50721	8853	(8749)
	I005I	2968	(3009)
	2000	902	(700-I000)
	I000	525	(-)
	500	300	(-)
	I00	77	(-)
	I0	9	(-)

С помощью формулы (2.40) мы предприняли попытку прогнозирования объема словаря для разных текстов из разных языков вплоть до объема $N = 10^7$ (см. табл. 2.I3). При сравнении данных следует учесть различие между жанрами (подъязыками) и различие в подсчете объема словаря (в словоформах или лексемах). Например, по данным научно-технических текстов английского языка (табл. 2.I3, е, ж) при $N = 10^7$ прогнозируется объем текста 36 000 - 38 000 словоформ, в то время как в смешанных английских текстах ожидаемый объем словаря при такой же длине текста - I48 000 словоформ (табл. 2.I2). При сравнении данных, взятых из разных языков, надо учесть и разницу в степени аналитизма языков; так, например, в текстах по электронике в английском языке прогнозируется 38 000 словоформ при $N = 10^7$, в то время как в русских текстах по электронике при такой же длине текста ожидаемый объем словаря - 94 000 словоформ. В агглютинативном казахском языке (в газетных текстах) ожидается объем словаря - 230 000 словоформ при $N = 10^7$.

\+ \+ \+

Подытоживая сказанное, можно констатировать следующее.

Наряду с чисто эмпирическим подходом к проблеме аналитического выражения связи между объемами словаря и текста в современной квантитативной лингвистике намечаются и некоторые направления теоретического анализа изучаемой зависимости по существу. При этом могут применяться как гипотетико-теоретический, так и гипотетико-эмпирический подходы (например, при выводе формул в гл. 2.3). Как и при анализе частотной структуры текста (гл. 2.2), так и в данном случае можно использовать разные модели для описания изучаемых явлений и разные формулы для аппроксимации эмпирических данных (распределений) в зависимости от принятых исходных постулатов.

Таблица 2.13

Наблюдаемый и ожидаемый объем (L или V) словаря в зависимости от длины текста (N) по данным разных языков по формуле: $L' \text{ (или } V') = Ne^{-a(\ln N)^b}$

а) Латышский язык – газеты, лексемы (Latviešu val., 1969)

N	L	L'
50000	7065	7025
100000	9834	9919
200000	13389	13510
300000	16103	15912
500000	-	19200
10^6	-	24000
10^7	-	37000

Параметры:

$a = 0{,}003736$

$B = 2{,}5304$

б) Чешский язык – технич. тексты, словоформы (Bečka, 1972)

N	V	V'
25000	4829	4827
75000	9603	9626
125000	13056	13050
175000	15858	15853
500000	-	28200
10^6	-	40000
10^7	-	114000

Параметры:

$a = 0{,}01123$

$B = 2{,}1539$

в) Казахский язык – газеты, словоформы (Ахабаев, 1971)

N	V	V'
25000	9088	9161
50000	15047	14875
100000	23895	23523
150000	29785	30378
500000	-	61000
10^6	-	87000
10^7	-	230000

Параметры:

$a = 0{,}001372$

$B = 2{,}8488$

г) Польский язык – А.Мицкевич, словоформы, (Sambor, 1970)

N	V	V'
12172	3434	3458
29787	6146	6044
48255	8026	7998
64510	9250	9398
100000	-	11800
500000	-	25000
10^6	-	33000
10^7	-	60000

Параметры:

$a = 0{,}00364$

$B = 2{,}6081$

д) Украинский язык – О.Довженко, словоформы (Дарчук, 1975)

N	V	V'
5000	1629	1629
10000	2637	2646
15000	3504	3482
20000	4195	4214
100000	-	11500
500000	-	28000
10^6	-	40000
10^7	-	110000

Параметры:

$a = 0{,}01055$

$B = 2{,}1783$

е) Английский язык – суд.механизмы, словоформы (Лукьяненков, Нешитой, 1975)

N	V	V'
50495	4871	4849
100970	6858	6882
201866	9470	9520
302156	11314	11360
403966	12975	12832
500000	-	14000
10^6	-	18000
10^7	-	36000

Параметры:

$a = 0{,}01235$

$B = 2{,}2019$

ж) Английский язык – электроника, словоформы (Алексеев, 1968)

N	V	V'
50000	5399	5437
100000	7853	7728
150000	9361	9371
200000	10582	10682
500000	-	15600
10^6	-	20000
10^7	-	38000

Параметры:

$a = 0{,}009152$

$B = 2{,}3057$

з) Румынский язык – электроника, словоформы (Ванн, 1966)

N	V	V'
50000	6785	6841
100000	10281	10070
150000	12477	12479
200000	14292	14454
500000	-	22400
10^6	-	30000
10^7	-	68000

Параметры:

$a = 0{,}008148$

$B = 2{,}3086$

и) Русский язык – электроника, словоформы (Калинина, 1968)

N	V	V'
50000	9464	9388
100000	14062	14168
150000	17263	17803
200000	21468	20818
500000	-	33000
10^6	-	45000
10^7	-	94000

Параметры:

$a = 0{,}004284$

$B = 2{,}5058$

3. ФОНЕТИЧЕСКИЙ, ГРАММАТИЧЕСКИЙ И СЕМАНТИЧЕСКИЙ АСПЕКТЫ ИССЛЕДОВАНИЯ ЛЕКСИКИ

Согласно теоретико-методологическим установкам, принятым в настоящей работе (см. гл. I.3), исследование лексики языка может производиться в разных лингвистических аспектах (на разных подуровнях лексики), в частности в аспектах взаимосвязи и взаимопроникновения лексики с фонетическим, грамматическим и семантическим уровнями языка. В соответствии с принципами квантитативно-системного подхода к изучению лексики основными методами анализа являются группировка (классификация) и моделирование с помощью распределений.

3.1. ФОНЕТИЧЕСКИЙ АСПЕКТ

Фонетическая классификация слов. Каждое слово, как элемент лексической системы языка, несет свою конкретную звуковую (фонемную, графемную) ф о р м у, по которой можно идентифицировать слово и включить его в определенную формационную, в данном случае лексико-фонетическую группу слов. Фонетическая классификация слов, по словам А.И. Смирницкого (1956, с. 140), "безусловно существенна для характеристики внешнего облика лексики данного языка". Она существенна также для типологического изучения языков и стилей. Несмотря на известную независимость плана выражения от плана содержания, фонетическая структура слова несет на себе отпечаток более высоких уровней языка, и, следовательно, изучение фонетики слова может играть немаловажную роль в познании более общих закономерностей функционирования слов в языке.

При квантитативно-системном исследовании лексико-фонетических групп ставится задача выявить особенности количественного распределения единиц (слов) по их фонетическому строению (по началу и концу слова, по фоно- и графемотактике, по моделям дистрибуции звуков, фонем или графем, по длине слова) с учетом взаимосвязи с другими подуровнями лексики.

Для иллюстрации приводим примеры из разных языков, причем в основном исходим из имеющихся данных о линейных последовательностях букв, учитывая то, что при автоматическом анализе текстов имеем дело с буквенными (графемными) цепями. Во многих случаях можно по данным буквенных цепей сде-

лать обоснованные выводы и о распределении звуков или фонем.

Знание особенностей строения слова с точки зрения его н а ч а л а и к о н ц а важно не только для изучения типологии языков, но и для решения вопросов автоматической обработки текста, в частности для разработки алгоритмов членения речи на значимые единицы и для морфологического анализа этих единиц.

Сравнение данных из разных языков показывает сходства и различия фонетического строения слова в отношении анлаута и ауслаута. Например, пять наиболее частотных букв в начале слова в эстонском и русском языках (по данным большого ортологического словаря эстонского языка и Акад. словаря русского языка в I7 томах, см. Андреев Н.Д., I967, с. 280) распределяются следующим образом (в процентах):

эстонский - k (I6,I), p (I0,2), s (9,2), t (8,2), v (6,3)
русский - п (I9,I), с (9,I), о (8,I), н (7,I), в (6,5)

Наряду с различиями в распределении начальных букв можно отметить и общие черты: в обоих случаях наблюдается одинаковая концентрация слов, начинающихся с пяти наиболее частотных букв; в эстонском словаре они составляют 50,0 %, в русском языке - 49,9 %. В общей сложности в обоих языках имеется большой перевес согласных элементов в анлауте - в эстонском языке 84,5 % согласных против I5,5 % гласных, в русском языке 83,7 % против I6,3 %.

Мы располагаем также данными о частотности букв в конце слова по данным словарей разных языков, например, в ауслауте эстонского слова преимущественно встречаются гласные e и a, а также согласные s, k, t . Однако эти данные характеризуют словарь лексем в "основной" форме (существительные и др. в именительном падеже ед. числа, глаголы в форме инфинитива). Для выяснения особенностей употребления слов в живой речи (в тексте) требуется провести исследование на основе текстов.

Распределение слов в т е к с т е в разных языках (Setälä V., I972; Лесохин М.М. и др., I982) на основе пяти наиболее частотных начальных букв:

эстонский - k (I4,I), t (9,5), s (8,8), m (8,3), p (7,4);
финский - j (I3,0), s (I0,9), k (9,9), h (9,5), t (9,3);
русский - в (II,7), п (I0,9), н (9,6), с (8,I), о (7,9).

Здесь бросается в глаза некоторое различие между близкородственными эстонским и финским языками. Но во всех рассматриваемых языках к наиболее частотным буквам (звукам) в анлауте относится /в/, а эстонскому и финскому языкам общими

являются еще /к/, /т/.

Соответствующие данные о конце словоформы в тексте:
эстонский - a (I9,I), e (I7,7), s (I3,0), d (II,6), i (II,5);
финский - n (28,I), a (23,7), ä (I3,I), i (II,0), e (I0,I);
русский - й (22,I), ь (22,0), е (I2,3), а (9,7), я (9,5).

Приведенные данные о распределении частот букв в начале и конце слова получены на основе анализа смешанных текстов. Известно, что при более подробном анализе можно обнаружить некоторые различия между конкретными распределениями частот фонетических единиц в текстах разных функциональных стилей.

Ранговые формы распределений словоформ по признаку начальных или конечных букв в словаре и тексте приблизительно подчиняются л о г а р и ф м и ч е с к о м у закону (Тулдава Ю.А., I986). Здесь, как и при исследовании независимых частот букв (звуков, фонем), а также в некоторых других случаях выявляется закономерность, согласно которой упорядоченные частоты "элементарных" единиц (например, звуков или звукосочетаний), как правило, подчиняются логарифмическому (или экспоненциальному) закону убывания частот (см. также Орлов Ю.К., I976, с. I85). Эмпирическое распределение частот таких единиц может быть аппроксимировано также уравнением окружности (Пиотровский Р.Г., I975, с. I09-III).

Фонотактические типы слов. Еще более важным признаком при выявлении типов слов является фоно- или графемотактика, т.е. условия сочетания звуков (фонем) или букв в той или другой позиции в слове. Подробные статистические данные, относящиеся к распределению частот появления букв (звуков) и их сочетаний в пределах слова и с учетом их позиции в слове по данным различных языков можно найти в книге Н.Д. Андреева (I967); по русскому языку в статьях: Белоногов Г.Г., Фролов Г.Д., I963; Денисов П.Н. и др., I978.

Известно, что буквенный (звуковой) состав концов слов, коррелирующий с грамматическими признаками слов, помогает определению классов слов и в конечном счете (вместе с другими признаками) автоматическому анализу текста (см., например, Белоногов Г.Г. и др., I983). Большую помощь оказывают при этом обратные словари, особенно обратные частотные словари словоформ.

Слово может иметь разные типы звуковой (фонемной, графемной) структуры, состоящей из соотношений согласных (C) и гласных (V). Такие обобщенные фонетико-структурные типы, опи-

сываемые на уровне классов C и V (в дальнейшем CV-структуры), могут быть изучены с квантитативной точки зрения в рамках анализа распределения формационных групп слов в языке. В типологических исследованиях фонетической структуры слова часто обращаются к т.н. к а н о н и ч е с к и м формам для того, чтобы можно было на более общей основе проводить сопоставительный анализ структуры слова разных языков. В канонических формах не учитывается признак длительности (краткость или долгота) звуков, и дифтонги рассматриваются как простые гласные элементы. При таком подходе можно, например, сопоставить последовательности наиболее частотных CV-структур односложных слов в разных языках (по данным словарей, см. Krámský J., 1966; Панкрац Г.Я., 1981):

русский - CVC, CCVC, CVCC, CCVCC, CV
немецкий - CVC, CVCC, CCVC, CCVCC, CV
английский - CVC, CVCC, CCVC, CV, CCVCC
казахский - CVC, VC, CVCC, CV, VCC
турецкий - CVC, CVCC, VC
венгерский - CVC, CVCC, VC
эстонский - CVC, CVCC, CV, CCVC, CVCCC

Можно констатировать близость индоевропейских языков, с одной стороны, и близость турецких и финно-угорских языков, с другой стороны. Примечательно, что наиболее частотной структурой односложных слов во всех рассматриваемых языков оказывается CVC.

В тексте распределение частот CV-структур может сильно отличаться от словарного распределения, например, в эстонском тексте наиболее частотными структурами являются (в порядке убывания частот): CV, CVC, VC, CVCC, V, которые вместе взятые покрывают 86,7 % текста (всего возможных структур односложной словоформы в эстонском языке - 15; подробнее см. Тулдава Ю.А., 1978).

По данным английского языка , кроме упомянутых структур односложных слов, наиболее частотными оказываются в словаре еще двухсложные CVCVC и CVCCVC (Слипченко Л.Д., 1973), а в тексте - CVCVC (Roberts A.H., 1965).

Интерес может представлять квантитативное определение степени симметричности системы фонетических структур слов. Система симметрична, если зеркальные структуры (CV и VC, CCV и VCC и т.д.) встречаются одинаково часто. Симметричность автоматически увеличивается за счет тех случаев, когда сами

структуры имеют симметричную форму (CVC, CCVCC, CVCVC и т.д.). По имеющимся данным такие "автосимметричные" структуры покрывают в немецком языке около 40 % всего словаря односложных слов, в эстонском языке - 43 % словаря и 35 % соответствующего текста. Продуктивность автосимметричных структур наблюдается и в других языках, например, в украинском языке (Перебийніс В.С., 1970). В то же время отмечается, что среди продуктивных моделей слов есть и ассимметричные цепочки звуков. Сбалансированное соотношение между симметрией и асимметрией справедливо считается важным принципом, лежащим в основе строения и функционирования элементов системы языка как экспрессивного средства общения, а также его развития (Муравицкая М.П., Слипченко Л.Л., 1982, с.77).

Связь между фонетическими структурами начала и конца слова можно наглядно представить в виде т.н. параллелограмма Менцерата (Menzerath P., 1954) или в виде особой матрицы - таблицы сопряженности. Такие таблицы нами составлены по имеющимся данным о распределении односложных слов в немецком и эстонском языках (данные по славянским языкам см. Заплаткина Н.И., 1975 и 1982). Таблицы содержат различную информацию о квантитативных свойствах фонетического (канонического) строения односложных слов с гласным ядром (см. табл. 3.1). Например, в немецком словаре односложное слово наиболее часто оканчивается на один согласный звук (50,3 %), то же самое в эстонском словаре (48,1 %), но в эстонском тексте преобладает нулевой согласный звук в конце односложного слова, т.е. слово оканчивается наиболее часто на гласный элемент ядра (45,2 %). Можно констатировать, что симметричность системы в немецком языке (в словаре) несколько выше, чем в эстонском, например, нет большой разницы между частотами зеркальных структур CVCC и CCVC (22,1 % и 18,1 %; ср. в эстонском языке 37,4 % и 4,2 %). В немецком языке достаточно близки и распределения сумм ($\sum$) строк и столбцов. Различия между языками объясняются типологическими особенностями языков, а в данном случае в некоторой степени также различиями в подсчете единиц.+

+ Таблица данных по немецкому языку составлена на основе материалов исследования П.Менцерата, причем за основу взяты словарные единицы - односложные лексемы (общий объем - 2225 единиц). В эстонском языке учитываются также возможные односложные флективные формы слов (общий объем словаря - 3293 единицы).

Таблица 3.I

Распределение канонических структур односложных слов по началу и концу слова (согласное окружение гласного ядра)

а) Словарь немецкого языка

Число согласных		В конце слова ∅	I	2	> 2	∑ (%)
В	∅	0,6	2,I	2,5	0,6	5,8
начале	I	4,4	28,7	22,I	4,9	60,I
слова	2	I,9	I8,I	I0,2	I,6	3I,8
	>2	0,4	I,4	0,5	0,04	2,3
	∑ (%)	7,3	50,3	35,3	7,I	I00,0

б) Словарь эстонского языка

Число согласных		В конце слова ∅	I	2	3	∑ (%)
В	∅	0,4	2,9	2,7	0,03	6,0
начале	I	5,3	40,8	37,4	3,0	86,5
слова	2	0,3	4,2	2,I	0,5	7,I
	3	0,03	0,2	0,I5	0	0,4
	∑ (%)	6,0	48,I	42,4	3,5	I00,0

в) Текст эстонского языка

Число согласных		В конце слова ∅	I	2	3	∑ (%)
В	∅	5,6	II,9	2,8	0	20,3
начале	I	39,5	29,7	I0,0	0,06	79,3
слова	2	0,06	0,2	0,06	0,03	0,4
	3	0	0,03	0	0	0,03
	∑ (%)	45,2	4I,8	I2,9	0,I	I00,0

Обобщающую математическую оценку связи между двумя признаками - звуковыми структурами начала и конца слова можно выразить с помощью коэффициента взаимной сопряженности, который вычисляется по следующей формуле (Митропольский А.К., 1971, с. 346):

$$\Phi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1, \qquad (3.1)$$

где n_{ij} - частоты в клетках таблицы, $n_{i.}$ и $n_{.j}$ - суммы частот по строкам и столбцам соответственно, r - число строк, k - число столбцов. В качестве показателя силы связи между признаками (началом и концом слов) удобнее всего использовать квадратный корень из Φ^2, т.е. Φ. По данным немецкого словаря Φ = 0,141, по эстонскому словарю Φ = 0,102. Интересно отметить, что сопряженность между началом и концом односложного слова в тексте оказывается сильнее, чем в словаре (по эстонскому тексту Φ = 0,239). По-видимому, это связано с закономерностями ритмического членения потока речи.

Длина слова. Важным количественным показателем структуры текста и его словаря является длина слова и распределение слов по длине в тексте и в соответствующем словаре. Длина слова считается важным квантитативно-типологическим критерием, который прогнозирует не только структурные черты языка, но и индивидуальные и функциональные особенности текстов и словарей (см. Алексеев П.М., 1986). Данные о длине слова и о распределении слов по длине используют в наши дни также при решении задач автоматической переработки текстов (Вертель В.А., Вертель Е.В., 1970). Практически важным является вопрос об аналитическом описании распределения слов по длине и о связи длины слова с другими структурными характеристиками текста и словаря (частотность, словообразовательная активность, возраст слов и т.д.). С точки зрения квантитативно-системного подхода к исследованию лексики длина слова представляет собой системообразующий фактор, определяющий условия разбиения лексики на группы слов разной формационной структуры, и распределение слов по длине указывает на статистическую взаимосвязь между словами разной длины в процессе порождения речи.

При изучении длины слова в тексте за основу берется словоупотребление в виде словоформы. Длина слова в словаре зависит от объема словаря и от выбранной единицы подсчета -

отдельной словоформы или лексемы (слова в "основной" форме). Наиболее информативным является сопоставительный анализ длины словоформы в тексте и в соответствующем словаре.

Длину слова (словоформы) можно измерять в буквах, звуках, фонемах, в слогах или морфемах, в зависимости от возможностей и от поставленных задач.

Измерение длины слова в буквах (графемах) удобно в том смысле, что процедуру измерения можно легко автоматизировать. Важно то, что во многих языках системы букв и звуков (или фонем) и тем самым результаты измерения длины слова сильно коррелируют.

Опыт показывает, что средняя длина словоформы (СФ) и распределение СФ по длине в т е к с т е и в с л о в а р е существенно различаются. Например, по данным ЧС английского языка (Kučera H., Francis W.N., 1967, с. 365-366) средняя длина СФ в тексте объемом N ≈ I милл. словоупотреблений (СУ) составляет 4,74 буквы, а в соответствующем словаре (объемом V ≈ I00 тыс. разных СФ) средняя длина СФ равняется 8,I3 буквам. Причем короткие СФ (1 - 4 буквы) покрывают 34,5 % текста, в то время как в словаре они составляют только 4 %. В некоторых языках (гл. обр. флективно-синтетического и агглютинативного строя) в распределении СФ в тексте обнаруживается двухвершинность (бимодальность) с пиками, например, на двух- и четырехбуквенных (-звуковых) СФ. Это объясняется тем обстоятельством, что распределение СФ неоднородно: в нем участвуют короткие служебные слова и более длинные знаменательные слова. Распределение СФ по длине в соответствующем словаре имеет, как правило, более "регулярную" форму, в частности отсутствует двухвершинность (по данным эстонского языка см. рис. 3.I; подробнее см. Тулдава Ю.А., I986).

Следует отметить, что как средняя длина СФ, так и распределение СФ по длине существенно варьируются в текстах (и словарях) разных подъязыков. Так, например, средняя длина СФ в эстонском тексте колеблется между 4,8 (в речи персонажей художественной прозы) и 7,I буквами (в научно-техническом тексте). Доля коротких СФ (2-3 буквы) в речи персонажей достигает 3I %, в то время как в научно-техническом тексте их только I5 %. (Аналогичные данные по другим языкам приводятся в работах: Андреев Н.Д., I967, с. 247; Никонов В.А., I978, с. I07; Папп Ф., I980, с. 22; и др.).

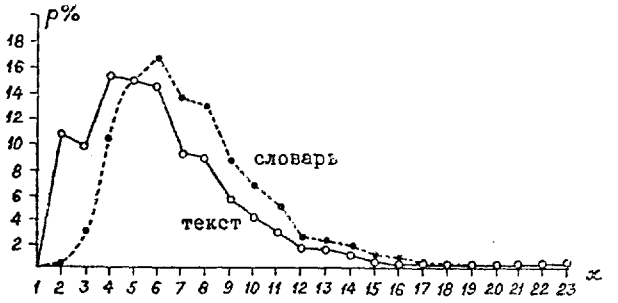


Рис. 3.1. Распределение словоформ по длине в тексте и в соответствующем словаре эстонского языка (x - длина в буквах, p % - доля в процентах)

Естественно, что длина СФ варьируется также по частям речи и по частотным зонам (см. Алексеев П.М., 1986).

Опытным путем установлено, что распределение СФ по длине (измеренной в буквах или звуках) в словаре хорошо аппроксимируется л о г н о р м а л ь н ы м распределением (Herdan G., 1966; см. также Тулдава Ю.А., 1986, с. 151). Считается, что логнормальность распределения СФ по длине отвечает принципу оптимального кодирования информации и отражает "стремление к ясному и безошибочному различению слов" (Herdan G., 1966, с. 205). По существу, логнормальное распределение означает, что выбор слова определенной длины зависит в некоторой степени от длины предшествующего слова, благодаря чему в тексте чередуются длинные, короткие и средние по длине слова (см. Пиотровский Р.Г. и др., 1977, с. 204). Однако распределение СФ по длине в тексте, в частности при двухвершинности распределения, требует иного подхода к аналитическому описанию распределения. Г.Я. Мартыненко (1965) предлагает в таком случае рассматривать отдельно служебные и знаменательные слова и показывает, что при их раздельном рассмотрении удается хорошо аппроксимировать эмпирические данные комбинированной показательно-степенной функцией. Можно также рассматривать распределение в виде накопленных частот, и тогда подходит теоретический закон Вейбулла.

Измерение длины слова в с л о г а х (или морфемах) имеет особое значение при выявлении т.н. глубины слова. Такие данные используются в педагогике, психолингвистике, стилистике, а также в типологических исследованиях языков.

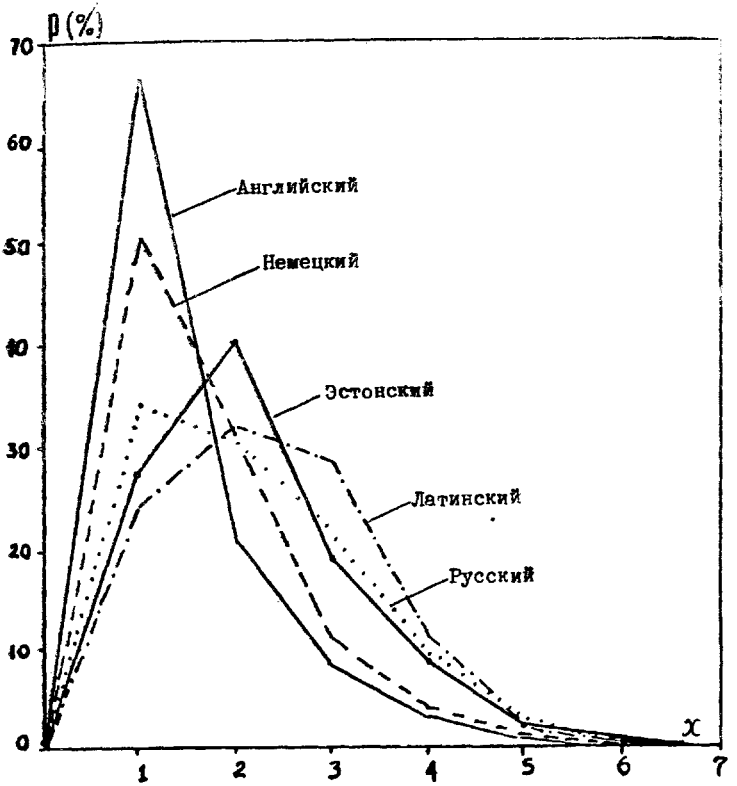


Рис. 3.2. Распределение словоформ по длине в тексте разных языков (x – средняя длина в слогах, p (%) – покрываемость текста в процентах)

Особенно ярко проявляется дифференциация языков при сравнении распределений СФ по длине в тексте (при идентичности функциональных стилей; см. табл. 3.2). Оказывается, например, что односложные СФ составляют в английском художественном тексте 66,8 %, в турецком тексте такого же стиля – только 18,8 %. В эстонском тексте односложные СФ покрывают от 25 до 40 % в зависимости от подъязыка и стиля (в художественной прозе в среднем 28,8 %). В русском языке односложные словоформы покрывают около 33 % художественного текста. Сходства и различия в структуре СФ разных языков особенно ясно выступают при их сопоставлении на графике (рис. 3.2).

Распределение СФ по длине, измеряемой в слогах, хорошо описывается логнормальным распределением, так же как и при измерении длины СФ в буквах. Поэтому можно принять логнормальный

Таблица 3.2

Распределение словоформ (СФ) по длине в тексте по данным выборок из художественной прозы разных языков. (Данные взяты из работ: Fucks W., 1956; Herdan G., 1966; Zsilka T., 1974; Якубайтис Т., 1963.)

Кол-во слогов	Английский	Французский	Немецкий	Румынский	Венгерский	Русский яз.	Латышский	Эстонский	Латинский	Турецкий
I	66,8	55,8	5I,7	45,5	34,4	33,9	33,8	28,8	24,2	I8,8
2	2I,I	27,9	3I,6	28,4	30,4	30,3	38,4	40,2	32,I	37,8
3	8,2	I2,9	II,I	I8,6	20,7	2I,4	I9,4	I8,5	28,7	37,0
4	3,3	2,9	4,3	6,7	9,9	9,7	7,0	8,8	II,6	I2,I
5	0,5	0,5	I,0	0,7	3,4	3,5	I,2	2,5	2,8	3,6
более 5	0,I	0,0	0,3	0,I	I,2	I,2	0,2	I,2	0,6	0,7
Ср. длина СФ (в слогах)	I,50	I,64	I,82	I,89	2,22	2,23	2,04	2,20	2,39	2,46

Таблица 3.3

Комплексное распределение словоформ (СФ) по длине в звуках и слогах по выборке из художественной прозы эстонского языка

Звуков / Слогов	2	3	4	5	6	7	8	9	I0	II	I2	I3	I4	I5	Число СФ	Число слогов	Число звуков	Ср. длина в звуках: СФ	Ср. длина в звуках: слога
I	II8	86	53	4											26I	26I	7I7	2,75	2,75
2		46	I08	I50	77	I9	2								402	804	I93I	4,80	2,40
3				I0	42	8I	44	I8	I						I96	588	I393	7,II	2,37
4						4	24	4I	20	9	3				I0I	404	924	9,I5	2,29
5						I		I	4	5	I0	2	4	I	28	I40	328	II,7I	2,34
6										2	2	4	2	2	I2	72	I56	I3,00	2,I7
Всего	II8	I32	I6I	I64	II9	I05	70	60	25	I6	I5	6	6	3	I000	2269	5449	5,45	2,40

закон за основу. Однако существуют и другие возможности. Распределение СФ по слоговой длине часто аппроксимируется функцией Чебанова-Фукса (Чебанов С.Г., 1947; Fucks W., 1956; Фукс В., 1957). Речь идет о модифицированном распределении Пуассона, при котором предполагается, что процесс формирования слов является случайным. Исследования показали, что еще лучшее соответствие эмпирических данных теоретическим достигается с помощью обобщенной формулы Гачечиладзе-Цилосани (модификация формулы Чебанова-Фукса), которая особым образом учитывает взаимодействие случайных и детерминированных процессов в речи (Гачечиладзе Т.Г., Цилосани Т.П., 1971).

В исследованиях по длине слова замечено, что существует закономерная связь между числом слогов (или морфем) и числом звуков (фонем) в слове. С удлинением слова (словоформы) отношение числа звуков к числу слогов уменьшается. Иначе говоря – чем длиннее слово, тем короче слоги в нем, или в более общем виде: "чем длиннее лингвистический конструкт, тем короче его компоненты" (з а к о н М е н ц е р а т а; см. Altmann G., 1980).

На рассматриваемом эмпирическом материале такая тенденция подтверждается. Например, по данным выборки из эстонской художественной прозы (см. табл. 3.3) наблюдается в общем уменьшение средней длины слога в связи с увеличением длины СФ: средняя длина слога односложных СФ равняется 2,75 звукам, двухсложных СФ – 2,40, трехсложных – 2,37 и т.д. вплоть до 2,17 в шестисложных СФ (отклонение от общей тенденции пятисложных СФ объясняется случайными факторами). Зависимость длины слога от длины слова приблизительно описывается логарифмическим законом

$$y = a + b \ln x , \qquad (3.2)$$

где y – длина слога в звуках, x – длина слова в слогах, a и b – параметры (в настоящем случае a = 2,75 и b = = – 0,33)[+]. Данная зависимость имеет своей основой закон "пропорционально убывающего относительного роста" (Ланд К.Ч., 1977, с. 388-389), который можно выразить дифференциальным уравнением

$$\frac{dy/y}{dx/x} = \frac{b}{y} . \qquad (3.3)$$

[+] Если x – длина слова в звуках (см. табл. 3.3), то изменяются только конкретные значения параметров a и b.

Это означает, что соотношение относительных приростов y и x (длин слога и слова) обратно пропорционально текущему значению средней величины y (длины слога). Тем самым получается характерное для логарифмической кривой замедляющееся изменение (возрастание или убывание) y по мере увеличения x.

Обобщающую и содержательно интерпретируемую модель закона Менцерата, т.е. модель для описания связи между "конструктом" и его компонентами предлагает Г. Альтманн (Altmann G., 1980). Модель представляется в виде системы уравнений, объединенных под общей формой:

$$y = ax^{b}e^{-cx}, \qquad (3.4)$$

где y - длина компонента, x - длина конструкта; a, b и c - параметры. Если $b = 0$, получается экспоненциальная функция $y = ae^{-cx}$; если же $c = 0$, то имеем степенной закон $y = ax^{b}$. Модель интерпретируется с помощью содержательных постулатов и дифференциальных уравнений.

Известно, что длина слова прямо или косвенно связана со многими структурными характеристиками словаря и текста. Рассмотрим связь между д л и н о й слова и ч а с т о т о й его появления в тексте. Тот факт, что короткие слова в среднем более частотны, чем длинные слова, свидетельствует, по-видимому, о действии принципа экономии в процессе коммуникации, о котором говорил еще Дж. Ципф (Zipf G.K., 1935). Основываясь на этих предпосылках, французский исследователь П. Гиро (Guiraud P., 1954) сформулировал закономерность, согласно которой для каждого языка можно определить константу (C), указывающую на связь между длиной слова (x) и его рангом (i) в частотном списке:

$$C = \frac{\lg 2i}{x}. \qquad (3.5)$$

Эта формула достаточно хорошо описывает упомянутую связь, причем для английского языка определяется $C \approx 0{,}59$.

По сути дела, формула (3.5) выражает логарифмический закон связи между длиной и частотой слова. Такую связь можно вывести из закона Ципфа-Мандельброта. Действительно, точный вывод из закона Ципфа-Мандельброта показывает, что длина слова (x) должна быть пропорциональна логарифму его вероятности, т.е. линейно зависеть от логарифма его ранга (i) (Калинин В.М., 1964):

$$x_i = a + b \ln i, \qquad (3.6)$$

где a и b – константы. Эта более точная (и теоретически более обоснованная) формула еще лучше описывает связь между длиной и частотой словоформы по данным многих экспериментов.

Другой возможностью является аналитическое описание связи между "накопленной" средней длиной слова

$$X_i = \frac{1}{i} \sum_{1}^{i} x_i$$

и его рангом (i). Предполагая и здесь логарифмическую зависимость между X_i и i, мы можем вычислить X_i по формуле:

$$X_i = \alpha + \beta \ln i, \qquad (3.7)$$

где α и β – параметры. Соответствие эмпирических и теоретических величин вполне хорошее (Тулдава Ю.А., 1986, с. 156). Обычно считается, что данная зависимость наилучшим образом аппроксимируется распределением В е й б у л л а, которое впервые было применено при решении этой задачи Г.Г. Белоноговым (1962). В наших обозначениях формула принимает вид

$$X_i = X_\Pi \left(1 - e^{-ci^k}\right), \qquad (3.8)$$

где X_Π – предел средней длины слова (в данной совокупности), i – ранг слова, c и k – параметры.

Зависимость между длиной слова и его частотой является взаимной, хотя можно утверждать, что решающую роль играет здесь частота употребления. Уже давно было замечено, что если частотность слова возрастает, то его форма подвергается редукции (см., например, Мартине А., 1963). Этим объясняются и многочисленные случаи сокращений и усечений в современных языках, например, когда отсекается либо начало, либо конец слова: кино(театр), ретро(спективный стиль); эст. kopter < helikopter 'вертолет'; англ. (omni)bus 'автобус'.

3.2. ГРАММАТИЧЕСКИЙ АСПЕКТ

Лексика и грамматика. С лексикологическим изучением слова тесно связан грамматический аспект анализа, так как "в реальной истории языка грамматические и лексические формы и значения органически связаны" и "всякое слово оформлено уже тем, что оно несет известные грамматические функции,

занимает определенное место в грамматической системе языка" (Виноградов В.В., 1947, с. 7 и след.) В рамках квантитативно-системного исследования лексики нас интересует возможность классификации лексики на основе грамматических признаков и выявление закономерностей распределения полученных классов (лексических групп) в словаре и тексте. В данной работе мы ограничимся изучением некоторых аспектов *словообразования*, позволяющих выделить лексико-формационные группы по признаку морфемного строения слова, и рассмотрением лексико-грамматических классов слов - *частей речи*, определяемых в рамках грамматики с помощью синтактико-морфологических критериев. Выделение основных словообразовательных групп и лексико-грамматических классов слов, а также квантитативный анализ их структур и условий распределения в конкретных языках (или подъязыках) имеет значение первостепенной важности для характеристики языков (подъязыков) в типологическом плане и способствует решению многих актуальных проблем прикладного языкознания.

Названные проблемы и задачи исследования относятся к "пограничной" области между лексикологией и грамматикой. Хотя некоторые исследователи причисляют словообразование к лексикологии (А.И. Смирницкий, К.А. Левковская), в наши дни все же преобладает мнение, что словообразование относится к грамматике или составляет самостоятельный раздел науки о языке (обзор см. Немченко В.Н., 1984, с. 9). При этом всегда подчеркивается, что словообразование тесно связано с лексикологией. Такие разделы грамматики, как морфология и синтаксис связаны с лексикологией в первую очередь тем, что они могут служить базой для стратификации лексики. Рассматривая в данной работе части речи как основные представители синтактико-морфологического уровня, мы не исключаем возможности анализа и других лексико-грамматических групп слов, имеющих значение для лексикологического исследования конкретного языка. В этой связи можно упомянуть исследования по "классификаторной морфологии" (Viks Ü., 1980), которые в качестве дополнительного результата дают классификацию лексики на основе формообразования ("флективной" морфологии).

Словообразовательная структура слов. Наиболее общим способом классификации слов по их словообразовательной (морфемной) структуре является их распределение по основным структурным типам: *корневым*, *производным* и *сложным* словам (с подтипами сложно-производных,

сложносокращенных, "сверхсложных" слов и др.; см. Ахманова О.С., 1966, с. 422). По имеющимся данным, в нормативном ортологическом словаре эстонского языка (объемом 115 000 слов) словообразовательные типы слов распределяются следующим образом:

корневые слова	– 5 %	несложные слова – 40 %
производные несложные	– 35 %	
производные сложные	– 25 %	сложные слова – 60 %
непроизводные сложные	– 35 %	

Трудно сравнивать квантитативные характеристики разных языков, если нет точных указаний на распределение подвидов словообразовательных типов слов и на вид и объем выборки. Например, Л. Хакулинен (1953, с. 288) приводит следующие данные по финскому языку: корневых слов 12 %, производных слов 44 % и сложных слов 44 %. При этом неизвестно, как распределяются сложные и несложные производные слова и на каком материале произведено исследование. Ф. Папп (1980, с. 23) более точен: он указывает, что среди 31 000 существительных академического словаря венгерского языка 55 % сложных слов (с суффиксами или без них); 42 % существительных не содержат в себе суффикса (при этом они могут быть простые или сложные). Общее число исконных корневых слов в венгерском словаре около 6000, т.е. 10,3 % при объеме словаря 58 000 лексем (Папп Ф., 1969).

В русском языке по данным А.Н. Тихонова (1983) непроизводные корневые слова составляют около 13 % ((на материале разных словарей общим объемом 145 000 слов), остальные 87 % являются производными в широком смысле (т.е. включающими как аффиксальные, так и сложные слова). Доля сложных слов в русском языке, по-видимому, не превышает 8 % (по данным большого украинско-русского словаря, насчитывающего 120 000 слов; см. Клименко Н.Ф., 1974).

В словаре газетной лексики современного немецкого языка типы слов распределяются следующим образом: корневых слов 5 %, несложных производных слов 9 %, сложных и сложно-производных слов 83 %, сокращений 3 % (Harlass G., Vater H., 1974). В английском языке (подъязык не указан) среди существительных 18 % корневых, 67 % производных (аффиксальных и неаффиксальных) и 15 % сложных слов (Ginzburg R.S. et al., 1966).

Можно сделать вывод, что во всех рассматриваемых языках (финно-угорских и индо-европейских) корневые слова составля-

ют небольшую долю словаря (от 5 до I8 %), в то время как различие между языками проявляется прежде всего в распределении сложных слов: 45 – 60 % в финно-угорских языках, около 80 % в немецком языке, I5 % в английском языке и менее I0 % в русском языке.

В т е к с т е, т.е. при употреблении в речи, обнаруживаются совсем другие соотношения. Например, в немецком и английском языках (в газетных текстах) корневые слова составляют 70 и 75 % соответственно, производные слова I8 и 23 %, сложные слова – I2 и 2 % (Кубрякова Е.С., I970). В эстонском языке простые корневые слова покрывают 70 – 80 %, производные и сложные слова вместе взятые 20 – 30 % текста (колебания в зависимости от подъязыка). Как известно, на основании данных о частотности употребления в тексте разных морфологических, в т.ч. словообразовательных типов слов была построена типологическая классификация языков Дж. Гринберга (I963). Вместо процентов Дж. Гринберг применил особые индексы, выражающие отношение количества определенного типа морфем к количеству слов текста (словоупотреблений). Например, индекс словосложения (compositional index) указывает на отношение корневых морфем (R) к числу слов текста (W). По уточненным данным Е.С. Кубряковой (I970) индекс R/W составляет для газетных текстов германских языков: английский – I,02; голландский и датский – I,II; немецкий – I,I2; шведский – I,I3; исландский – I,I4. В эстонском газетном тексте индекс принимает значение I,20; в русском газетном тексте – I,04 (т.е. в среднем I04 корневых морфемы на I00 слов текста, иначе говоря, 4 сложных слова на I00 слов текста, если считать, что сложное слово русского языка обычно состоит из двух корней). Отношение общего числа морфем (корневых, словообразовательных и словоизменительных) к числу словоупотреблений (M/W) по-своему измеряет степень синтетизма языка (ср. гл. 2.I). Очевидно, что показатели этого индекса для аналитических языков будут низкими, для синтетических – более высокими. В германских языках значения индекса распределяются следующим образом (Кубрякова Е.С., I970, с. I6I): английский – I,43; голландский – I,8I; датский – I,98; немецкий – 2,02; исландский – 2,09; шведский – 2,I3. В эстонском языке (на материале газетных текстов) индекс достигает значения 2,35. Еще выше значение индекса в таких языках, как санскрит, суахили, эскимосский (см. Гринберг Дж., I963).

Особого рода классификация лексики получается исходя

из словообразовательных классов слов (Белоногов Г.Г. и др., 1985). Класс определяется списком суффиксов и сочетаний суффиксов, совместимых со словообразовательной основой слова.[+] В русском языке обнаружено около 1250 словообразовательных классов, причем 10 наиболее частотных классов охватывают 60% всех слов русского языка. Здесь проявляется известный принцип концентрации и рассеяния единиц, который приводит к образованию "ядра" и "периферии" в распределениях языковых объектов. К наиболее частотным классам относятся (класс характеризуется словом-представителем и перечнем суффиксов; "Ø" обозначает нулевой суффикс):

масштаб: Ø, -н- (т.е. слова типа: масштаб, масштабный);
слаб-ый: Ø, -о-, -ост- (слабый, слабо, слабость);
порт : Ø, -ов- (порт, портовый).

В общей сложности в русском языке обнаружено около 10 тыс. словообразовательных основ, причем количество словообразовательных суффиксов и сочетаний суффиксов превосходит тысячу (Белоногов Г.Г. и др., 1984). С другой стороны, в русском языке зарегистрировано около 10 тыс. "моделей морфного строения" основы слова при общем числе 5000 морфов, в т.ч. 4500 корневых морфов, 425 отдельных суффиксов и 75 префиксов (Ефремова Т.Ф., 1968).

Распределение производных слов по их словообразовательным формантам, т.е. по префиксам и суффиксам, позволяет выявить наиболее продуктивные и употребительные типы слов.[++] Например, по данным Обратного словаря русского языка (Bielfeldt H.H., 1965) наиболее частотными суффиксами являются -ный/-ной (около 9800 слов типа: трудный, лесной), -ние/-ие (3200: пение, шествие), -ка/-очка (3000: речка, лампочка), -ский/-ской (2700: русский, тверской), -ость (2500), -ник (1200) и др. (ср. Арапов М.В., Херц М.М., 1974). В Большом немецко-русском словаре О.И. Москальской (по данным Б.И. Барткова, 1983) наиболее частотными префиксами в немецком словаре являются: ver- (1100 слов с этим префиксом), un- (885), be- (783), er- (308, ent- (270); наиболее частот-

[+] Словообразовательная основа слова определяется как начальная часть слова, получаемая путем отделения максимального числа суффиксов.

[++] Б.И. Бартков (1982) различает "диахронную продуктивность" (объем класса, т.е. частотность в большом словаре) и "синхронную продуктивность" (по данным словаря неологизмов). Под употребительностью мы понимаем частотность в тексте.

ными суффиксами являются: -ung (IO 000), -ig (3800), -er (3000), -isch (2I00), -keit (2000). В тексте же (на материале газетных текстов) слова по частотности формантов ранжируются следующим образом: be-, ver-, er-, ge-, ent- и -ung, -er, -isch, -lich, -ig.

В отличие от лексико-фонетических групп (см. гл. 3.I), где была обнаружена логарифмическая (или экспоненциальная) зависимость в ранговом распределении слов, распределение слов по частотности словообразовательных формантов, как правило, подчиняется степенному закону, т.е. закону Ципфа. М.В. Арапов (I974) показал, что в достаточно обширном словаре ранговое распределение частот слов с данным суффиксом приблизительно подчиняется закону Ципфа с поправкой Мандельброта, т.е. распределение может быть аппроксимировано функцией типа $p_i = k(i+B)^{-\gamma}$ (ср. формулу (2.II)), где p_i – относительная частота слова с данным формантом, i – ранг, k, γ и B – константы.

Отношение частотности в тексте (F_T) к частотности в соответствующем словаре (F_C) выражает функциональную значимость, или (относительную) ф у н к ц и о н а л ь н у ю н а г р у з к у (ФН) рассматриваемого типа производных слов. Например, по данным английского языка (на материале ЧС Кучеры-Фрэнсиса при $F \geqslant 5$; см. Пиквер А., I973) в приведенном фрагменте (табл. 3.4) наибольшую функциональную нагрузку имеют существительные с суффиксами -y, -ment и -ion (например, policy, government, action) и прилагательные с суффиксами -ent и -ic (different, economic). Отношение F_T/F_C выражает по существу среднюю частоту данного типа слов в тексте: чем больше это отношение, тем больше повторяются слова с данным формантом в тексте, но тем меньше их относительное разнообразие. Существительные с суффиксом -er при низкой относительной функциональной нагрузке в тексте (ФН = 6,5) характеризуются множеством разных производных. Обратное соотношение (F_C/F_T) выражает, таким образом, степень разнообразия данного типа производных слов.

Квантитативное исследование может распространяться и на т.н. словообразовательные гнезда, под которыми понимаются группы однокорневых слов, объединенных отношениями словообразовательной производности. Обычно выделяется "вершина" гнезда как мотивирующее слово по отношению к остальным членам гнезда. Например, в английском языке выделяется гнездо с вершиной TIME 'время', которое характеризуется наибольшим

Таблица 3.4

Частотность и функциональная нагрузка словоформ с данным суффиксом в английском языке

№ пп.	Суффикс	Частота в тексте (F_T)	Частота в словаре (F_C)	Функциональная нагрузка ($F_T : F_C$)
1.	-ion	12772	794	16,1
2.	-ly (adv.)	10627	942	11,3
3.	-al (adj.)	7937	612	13,0
4.	-ate (verb)	7782	815	9,6
5.	-er (noun)	5221	803	6,5
6.	-ment	4991	296	16,9
7.	-ic	4691	396	15,9
8.	-y (noun)	3650	186	19,6
9.	-ent (adj.)	3486	192	18,2
10.	-ity	3196	350	9,1

деривационным и лексическим объемом: по суммарным данным нескольких английских словарей в нем содержится около 100 производных единиц (Беляева Т.М., Васильева Н.М., 1984). Способность слова быть производящей основой (как для аффиксальных производных, так и для сложных слов) называется словообразовательным потенциалом слова (СП). Среди простых, непроизводных слов русского языка высоким СП обладают такие имена существительные, как, например, вода (зарегистрировано 316 разных производных и сложных слов), свет (306), земля (216); глаголы бить (446), брать (393), делать (318); имена прилагательные белый (246), черный (236), старый (192) и др. (Тихонов А.Н., 1983).

На основе оценок СП можно судить о квантитативной структуре словообразовательных гнезд, можно классифицировать слова, например, разбить словарь на группы с высоким, средним и низким уровнем СП. Внутри гнезда можно выявить ядро и периферию по распределению частот употребления отдельных элементов гнезда.

Важно отметить, что по опытным данным наблюдается статистическая зависимость между оценками СП и частотностью производящих слов. Наиболее частотные слова обладают в среднем и наивысшими оценками СП, корреляция между частотностью вершины и мощностью словообразовательного гнезда статистически существенна (Бартков Б.И., 1983). При аналитическом изучении связи между частотностью и оценками СП обнаружива-

ется сложная зависимость, имеющая форму логисты, т.е. на достаточно обширном материале наблюдается медленное возрастание СП в связи с увеличением частотности, затем крутой подъем в зоне среднечастотных слов и стабилизация в зоне высокочастотных слов (см. Андрукович П.Ф., Королев Э.И., 1977). Результаты конкретного анализа словаря с точки зрения связи между частотностью и СП были на практике использованы при построении тезауруса и при оптимизации лингвистического обеспечения ряда автоматизированных информационных систем (Королев Э.И. и др., 1984).

Части речи. Объединение слов в лексико-грамматические классы, называемые частями речи, осуществляется на основе совокупного учета нескольких факторов: лексико-грамматического (категориального) значения, морфологических свойств, синтаксических функций. Несмотря на колебания в технике их выделения, части речи представляют собой принципиально стабильные категории, характеризуемые достаточно объективными показателями (Степанова М.Д., Хельбиг Г., 1978, с. 23). Важность изучения частей речи, в том числе их квантитативного анализа, подчеркивается тем обстоятельством, что в характере распределения частей речи в словаре и тексте раскрываются существенные типологические свойства данного языка, подъязыка или стиля.

Распределение частей речи в с л о в а р е зависит от объема словаря, причем наблюдается неуклонное нарастание доли существительных с увеличением объема словаря. Например, сравнивая словари лексем эстонского языка по данным выборок разного объема, можно констатировать, что процент существительных увеличивается от 44,2 до 61,6 и 75,0 (см. табл. 3.5). Обратную тенденцию можно наблюдать у всех остальных частей речи, кроме прилагательных, которые в данном случае сохраняют более или менее постоянную частотность в словаре (около II %). Доминирующее место имен существительных в большом словаре обусловливается частично тем, что имена существительные представляют собой главный источник обогащения словаря новыми словами.

Динамику изменения структуры словаря можно показать также на основе распределения частей речи по частотным зонам (табл. 3.6). Например, по данным ЧС лексем эстонского языка в зоне высокочастотных слов ($F \geq 10$) существительные составляют 35,2 % и глаголы – 24,5 %, но в зоне однократных

Таблица 3.5

Распределение частот частей речи (в %) в словаре эстонского языка по данным выборок разного объема: I – подвыборка, II – ЧС лексем, III – ортологический словарь (Oigekeelsussõnaraamat, 1976)

Часть речи	I	II	III
Существительное	44,2	61,6	75,0
Глагол	26,6	14,0	8,5
Прилагательное	9,5	11,7	11,4
Наречие	10,7	9,6	4,2
Местоимение	4,4	0,4	0,07
Числительное	1,3	1,0	0,2
Пред- и послелоги	2,2	1,0	0,1
Союз	0,9	0,2	0,03
Междометие	0,2	0,5	0,5
Всего (%)	100,0	100,0	100,0
Объем словаря	2200	14650	115000
Объем текста	5000	100000	-

слов (f = 1) их частоты равняются 65,8 и 11,1 %, соответственно. Существительные тяготеют к зонам редких слов, в то время как глаголы (а также наречия, местоимения и др.) являются преимущественно высоко- и среднечастотными. Прилагательные оказываются и здесь довольно стабильными. (Аналогичные данные по русскому языку см.: Jiráková I., 1976; по латышскому языку – Якубайтис Т.А., 1981).

Частотность частей речи в т е к с т е не зависит существенно от величины выборки, и тем самым их распределение в тексте является важным стиледифференцирующим фактором. Индивидуальные различия в распределении частей речи находятся в определенных границах, обусловленных функциональным стилем. Разница в употреблении частей речи обнаруживается особенно между художественными и нехудожественными стилями. Например, по данным ЧС русского языка (1977) частоты существительных ранжируются по жанрам: газеты – 32,8 %, научно-публицистические тексты – 31,0 %, художественная проза – 23,4 %, драматургия – 20,4 %. Частоты местоимений ранжируются в обратном порядке: драматургия – 16,2 %, худож. проза – 14,9 %, научно-публ. – 11,6 % и газеты – 10,0 %. Частотность глагола в драматургии – 20,9 %, в худож. прозе – 19,0 %, в газете – 14,5 % и в научно-публ. текстах – 13,5 %.

В рамках одного жанра можно констатировать глубокую взаимосвязь и взаимообусловленность в употреблении частей речи в тексте. Например, имеется существенная отрицательная корреляция между употреблением имен существительных и местоимений, т.е. эти части речи "конкурируют" между собой при

назывании (номинации) предметов и явлений. Отрицательная связь существует, как правило, и между частотами глаголов и прилагательных: это обусловливается стилистическим различием "вербальности" и "квалитативности" при описании предметов и явлений. Соотношения частот употребления частей речи (существительных и глаголов, прилагательных и глаголов и т.д) применяется в стилистике и психологии в качестве стилеметрических показателей "номинальности", "вербальности", или "действенности" и т.п. при исследований индивидуального стиля (см., например, Busemann A., 1948; Antosch F., 1969; Головин Б.Н., 1971; Тулдава Ю.А., 1976; Якубайтис Т.А., 1981).

При сравнении данных из разных языков обнаруживается, что при соблюдении требования однородности (одножанровости) текстов различия бывают не очень большими. Можно отметить доминирующее положение имен существительных в текстах художественной прозы в финно-угорских языках (эстонский, финский, венгерский) – в среднем 30 – 31 %, а также в латышском, русском и украинском языках – в среднем 28 – 29 % (см. табл. 3.7). Сравнение распределений частот частей речи производится часто с помощью обобщающей теоретико-информационной меры энтропии, вычисляемой по известной формуле Шеннона (Shannon C.E., 1948):

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i , \qquad (3.9)$$

где H – энтропия, p_i – относительные частоты, $\log_2$ – двоичный логарифм. Мера энтропии указывает на степень "неопределенности" при выборе слов той или иной части речи, иначе говоря, бóльшая оценка энтропии связана с большей равномерностью распределения единиц. По данным настоящего исследования, эстонский язык близок к финскому (H = 2,62 и H = 2,60, соответственно), но отличается, например, от русского языка (H = 2,85), в котором распределение частей речи в тексте, судя по оценке энтропии, более равномерно.

Что же касается вида распределения частей речи в тексте (по типу "однообъектного" распределения; см. гл. I.3), то можно сослаться на основательное исследование Т.А. Якубайтис (1981), в котором на основе 100 выборок по 1000 словоупотреблений выявлено, что при упомянутых условиях эксперимента во всех подъязыках латышского языка распределение частей речи как высокочастотных (например, существительных, глаголов), так и низкочастотных (например, предлогов) может быть аппроксимировано нормальной кривой.

Таблица 3.6

Динамика изменения распределения частот частей речи (в %) по частотным зонам в словаре по данным ЧС лексем эстонской художественной прозы. Общий объем словаря L = 14654 лексем; объем текста N = 99898 словоупотреблений.

Частота / Часть речи	L ≥10	9	8	7	6	5	4	3	2	1	Весь словарь
Существительное	35,2	40,6	43,3	37,9	48,5	50,6	45,5	54,1	57,3	65,8	61,6
Глагол	24,5	22,6	27,0	29,3	22,7	23,8	20,7	20,5	17,8	11,1	14,0
Прилагательное	11,6	15,1	14,2	12,6	13,3	12,5	15,8	12,8	11,7	13,1	11,7
Наречие	16,6	17,0	12,0	14,7	11,2	10,9	15,2	10,4	11,5	9,0	9,6
Местоимение	3,4	0	0	0	0	0,3	0,7	0,6	0,3	0,1	0,4
Числительное	1,3	0,9	1,4	1,0	2,2	0,6	0,6	0,6	0,8	0,3	1,6
Пред- и послелоги	3,6	2,3	1,4	3,0	1,7	1,3	0,8	0,9	0,4	0,1	1,0
Союз	1,4	0	0	0,5	0	0	0	0,1	0	0	0,2
Междометие	0,2	0,9	0,7	1,0	0,4	0	0,7	0	0,4	0,5	0,5
Всего	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Таблица 3.7

Распределение частот частей речи (в %) в тексте художественной прозы разных языков (данные взяты из работ: Saukkonen P. et al., 1979; Zsilka T., 1973; Якубайтис Т.А., 1981; Клочкова З.А., 1968; Тищенко В., 1970)

Часть речи \ Язык	Эстонский	Финский	Венгерский	Латышский	Русский	Украинский
Существительное	31,7	29,7	30,0	28,1	28,7	29,2
Глагол	22,5	27,6	22,4	23,1	18,3	19,7
Прилагательное	6,0	7,8	10,0	5,4	7,9	6,8
Наречие	15,8	10,9	8,0	10,0	6,0	6,2
Местоимение	11,4	11,2	5,7	14,6	10,2	9,0
Числительное	1,1	1,2	1,5	1,2	1,2	1,1
Пред-, послелог	3,1	3,6	21,9	5,1	12,2	12,5
Союз	8,2	7,3		7,3	8,5	9,1
Междометие	0,2	0,7	0,5	0,6	0,0	0,0
Частицы	-	-	-	4,6	7,0	6,4
Всего (%)	100,0	100,0	100,0	100,0	100,0	100,0
Энтропия (H)	2,62	2,60	2,47	2,79	2,85	2,82

3.3. СЕМАНТИЧЕСКИЙ АСПЕКТ

Лексико-семантические группы. Одной из основных задач изучения лексики с точки зрения ее системных свойств является выявление лексико-семантических групп (ЛСГ). Это общее название всяких лексических групп, образовавшихся на лексико-семантическом уровне, т.е. на основе семантической связи слов. Но ЛСГ бывают разных типов и объемов.

По типу отношений между словами лексико-семантические группы подразделяются на группы с парадигматическими связями (сходство, противопоставленность, субординация) и синтагматическими связями (совместная встречаемость, сочетаемость). Благодаря парадигматическим отношениям слов обеспечивается возможность выбора единиц при порождении речи, и этим определяется важность исследования парадигматических ЛСГ как в качественном, так и в количественном плане (не уменьшая при этом роли синтагматического фактора, тесно взаимодействующего с парадигматическим).

К образованию ЛСГ можно подойти двояко. С одной стороны, возможен синтетический подход, когда образование группы начинается от элемента (от частного к общему). С другой стороны, при аналитическом подходе продвигаются от универсального множества к подмножествам (от общего к частному).

Исходной самостоятельной единицей лексико-семантического уровня следует считать лексико-семантический вариант (ЛСВ), с которого и начинается образование ЛСГ при синтетическом подходе. ЛСВ может быть основной единицей подсчета и при квантитативном исследовании ЛСГ, если удается точно выделить отдельные значения слова.

При а н а л и т и ч е с к о м подходе членение лексико-семантической системы языка происходит на основе системы дифференциальных семантических признаков (ДСП), которые отличаются друг от друга различной степенью обобщенности. При организации ДСП в систему может быть использован дихотомический принцип, при котором каждый признак представляется в виде бинарной привативной оппозиции как А/не-А, в которой один член характеризуется наличием, а другой - отсутствием признака (это можно выразить также через +/- или 1/0; последнее обозначение как бы определяет вероятность принадлежности элемента к множеству[+]). Вполне возможна и система при-

[+] Теоретически возможна и более сложная модель, в которой рассматривается вся шкала вероятностей от 0 до 1, или "степень принадлежности" элемента к данному множеству (согласно теории нечетких множеств).

знаков, где используются не только бинарные оппозиции, но и отношения большей размерности. Дихотомический принцип выбирается обычно по той причине, что бинарные оппозиции категориальных признаков типа нарицательность – ненарицательность, конкретность – неконкретность и др. имеют определенный содержательный (стилистический) смысл и оказываются удобными и наглядными при квантитативном анализе распределений слов. Вследствие того, что выделенные признаки являются неравнозначными, вполне естественно устанавливается иерархия (инклюзивность) в системе признаков и, следовательно, также в системе ЛСГ. М о д е л ь такой системы можно схематически представить в виде дихотомического дерева (см. рис. 3.3). Следует добавить, что на практике любая ветвь дихотомического дерева может быть продолжена или отсечена в зависимости от задач исследования. Как выясняется, одна и та же совокупность ДСП присуща определенному множеству слов, которые на этом основании объединяются в группы (ЛСГ) разных уровней и объемов. Как отдельные слова, так и ЛСГ, куда они входят, могут быть представлены в виде цепочек ДСП (например, нарицательное – конкретное – неодушевленное и т.д.) или с помощью цифрового двоичного кода – последовательностью единиц и нулей в порядке их следования по ветви дихотомического дерева. Последний способ удобен при процедурах автоматической классификации в сочетании с количественным подсчетом.

Для примера приводим количественные данные о распределении нарицательных имен существительных эстонского языка по принятой сокращенной схеме распределения ДСП (рис. 3.3). Конкретный материал взят из ЧС авторской речи эстонской художественной прозы (ХП) и из ЧС законченного художественного произведения (романа А.Х. Таммсааре "Правда и право") – отдельно из авторской речи (АР) и речи персонажей (РП) (подробнее см. Тулдава Ю.А., 1983а).

По бинарному категориальному признаку конкретность – неконкретность нарицательные имена подразделяются на две большие группы – конкретные и абстрактные существительные. По данным рассматриваемых словарей имена существительные распределяются следующим образом (в %):

	ХП	АР	РП
Конкретные	60,9	55,2	65,6
Абстрактные	39,1	44,8	34,4

Конкретные существительные преобладают во всех рассматриваемых словарях, но можно отметить статистически сущест-

венные различия между "нормой" (сводным словарем – ХП) и индивидуальным словарем авторской речи (АР), а также между словарями авторской речи и речи персонажей одного и того же художественного произведения (АР и РП).

В соответствующих текстах доля конкретных существительных оказывается еще выше:

	ХП	АР	РП
Конкретные	63,2	65,6	70,8
Абстрактные	36,8	34,4	29,2

Конкретные существительные покрывают, таким образом, примерно 60 – 70 % эстонского художественного текста, причем наибольший процент конкретных существительных (70,8 %) обнаруживается в речи персонажей. Преобладание конкретных существительных в этих текстах объясняется своеобразием художественного стиля по сравнению с другими функциональными стилями. Это подтверждается и исследованиями по другим языкам. Например, по данным немецких текстов доля конкретных существительных в художественной прозе – 73 %, а в научно-техническом тексте только 28 % (Кульгав М.П., 1971, с 18–19). Отмечается, что абстрактные существительные преобладают обычно там, где требуется обобщенность или сжатость изложения.

При дальнейшем членении ЛСГ к о н к р е т н ы х существительных можно использовать категориально-семантические признаки меньшей степени обобщенности (одушевленность – неодушевленность, лицо – нелицо и т.д.; см. рис. 3.3). Согласно этой схеме среди конкретных существительных выявляются четыре основных подгруппы: ЛСГ названий людей (Н I I I), ЛСГ названий животных (Н I I О), ЛСГ названий природных предметов (Н I О I) и ЛСГ названий артефактов (Н I О О). Их общее частотное распределение (в %) в словаре и тексте эстонской художественной прозы приводится в табл. 3.8.

Таблица 3.8

Распределение частот подгрупп конкретных существительных

ЛСГ (тип)	Словарь			Текст		
	ХП	АР	РП	ХП	АР	РП
Н I I I	23	21	24	23	29	35
Н I I О	4	6	9	3	6	7
Н I О I	25	26	25	33	33	28
Н I О О	48	47	42	41	32	30
Всего (%)	100	100	100	100	100	100

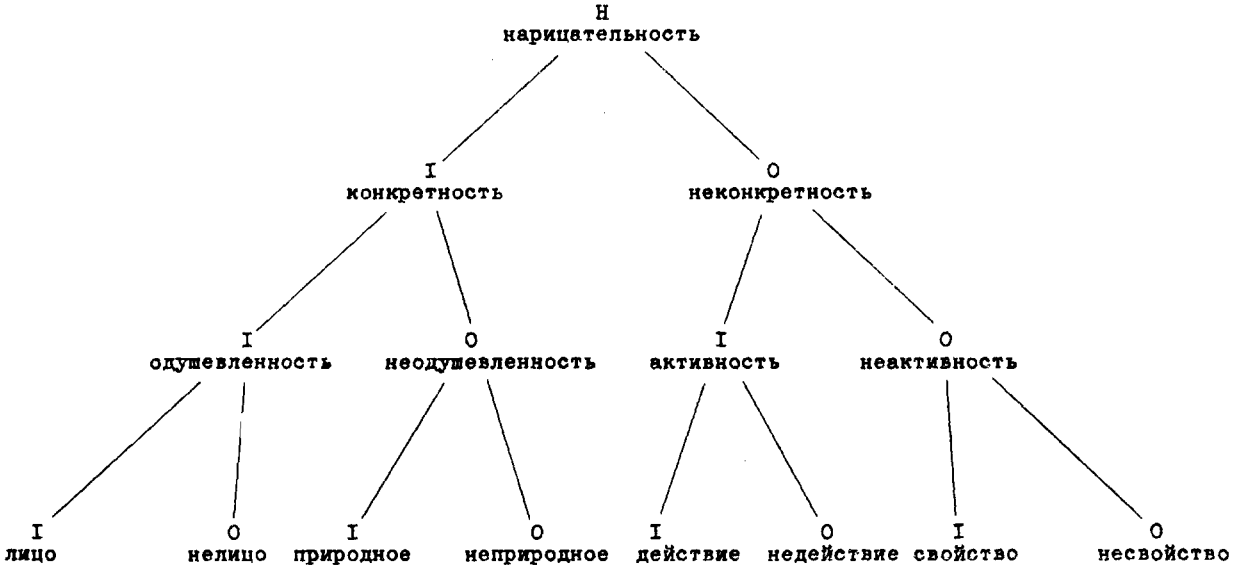


Рис. 3.3. Дихотомическое дерево дифференциальных семантических признаков (ДСП) для нарицательных имен существительных

Как видно из этого сопоставления, распределение частот ЛСГ конкретных существительных на категориально-семантическом уровне, хотя и показывает индивидуальные различия в определенных границах, в целом является довольно устойчивым и может считаться в общих чертах характерным для художественной прозы данного языка. Наибольшую частотность обнаруживает ЛСГ названий артефактов (Н I О О). Различия между распределениями отдельных стилей проявляются в тексте в большей мере, чем в словаре.

Разделение ЛСГ а б с т р а к т н ы х существительных на подгруппы проводится также на основе бинарных категориально-семантических признаков меньшей степени общности (активность – неактивность, действие – недействие и др.). На этом основании абстрактные существительные могут быть подразделены на четыре основные подгруппы – ЛСГ существительных, выражающих активное действие (по структуре Н О I I), ЛСГ существительных, выражающих состояние или результат действия (Н О I О), ЛСГ существительных, выражающих свойство (Н О О I) и ЛСГ прочих типов абстрактных существительных (Н О О О). Их частотное распределение в словаре и тексте эстонской художественной прозы (см. табл. 3.9) показывает, что, если не считать "остаточную" группу (Н О О О), в общей системе абстрактных существительных ЛСГ слов, выражающих активное действие (Н О I I), занимает первое место по частотности как в словаре, так и в тексте.

Таблица 3.9

Распределение частот подгрупп абстрактных существительных

ЛСГ (тип)	Словарь			Текст		
	ХП	АР	РП	ХП	АР	РП
Н О I I	37	43	27	22	25	I8
Н О I О	I0	I0	7	9	I0	6
Н О О I	5	8	7	5	5	8
Н О О О	48	39	59	64	60	68
Всего (%)	I00	I00	I00	I00	I00	I00

По частотному распределению подгрупп абстрактные существительные в несколько большей степени, чем конкретные существительные, являются стиледифференцирующими в художественной прозе, так как различия между стилями проявляются в распределении абстрактных существительных достаточно ясно не

только в тексте (как у конкретных существительных), но и в словаре, т.е. при расчленении конкретного инвентаря слов на лексико-семантические группы.

Полисемия. В современной лингвистике полисемия рассматривается как важная семантическая универсалия, которая может принимать различные формы в разных языках, но в основном подчиняется некоторым общим, в том числе количественным закономерностям. В последнее время появился ряд исследований, в которых затрагивается вопрос о системности полисемии в квантитативном плане (см., например, Поликарпов А.А., 1976 и 1987; Вишнякова С.М., 1976; Крылов Ю.К., Якубовская М.Д., 1977; Андрукович П.Ф., Королев З.И., 1977). Большой интерес к квантитативным закономерностям полисемии объясняется в наши дни главным образом практическими потребностями лексикографии, а также автоматической переработки текстов (информационный поиск, автоматический перевод). В то же время выявление квантитативно-системных характеристик полисемии может пролить новый свет и на общие закономерности функционирования лексико-семантической системы языка в целом. Сопоставительное исследование количественной стороны полисемии может дать дополнительные критерии для семантической типологии.

Исследуя полисемию как систему, целесообразно не отрывать собственно многозначность от однозначности; их следует рассматривать вместе как проявления одного и того же свойства слова - иметь одно или несколько значений. Однозначность выступает в таком случае как "нулевая степень" полисемии. Реализацией свойства слова - иметь некоторое число значений - является с е м а н т и ч е с к и й о б ъ е м слова, который подлежит количественному измерению. При этом мы исходим из положения, что многозначное слово (лексема) как единица системы языка представляет собой единое смысловое целое, объединяющее в плане содержания ряд виртуальных семантических вариантов, или отдельных "значений" слова. Существуют различные способы выявления смыслового содержания слова и разграничения отдельных его значений, и в данном вопросе наблюдается значительный разнобой в практической лексикографии. Опыт изучения полисемии показывает, однако, что в рамках одного какого-нибудь толкового словаря разграничение значений слов проводится достаточно последовательно, о чем свидетельствуют удачные попытки изучения общих квантитативных закономерностей полисемии на основе таких словарей (см. Папп Ф.,

1969; Крылов Ю.К., Якубовская М.Д., 1977).

Опытные данные говорят о том, что между показателями семантического объема разных частей речи имеются существенные различия. Например, по данным исследования Ф. Паппа (1967), в полном толковом словаре венгерского языка среднее количество значений у глагола – 2,3; у прилагательных – 1,9; у существительных – 1,6. В такой же последовательности располагаются части речи по данным тезауруса Роже английского языка (см. Вишнякова С.М., 1976): глаголы – в среднем 3,5 значений; прилагательные – 2,5; существительные – 2,1. Из полнозначных слов наименьший семантический объем имеет наречие (по данным английского языка – в среднем 1,4 значений). В толковом словаре эстонского языка наибольшее количество значений имеет также глагол (Тулдава Ю.А., 1979). Таким образом, в отношении этих языков подтверждается вывод о том, что "смысловой объем глагола шире, чем у имени" (Уфимцева А.А., 1968, с. 89).

Далее, можно констатировать, что семантический объем слова и количество слов с данным семантическим объемом в словаре статистически связаны: наибольшую долю в словаре составляют однозначные слова, затем следуют в порядке убывания частот слова с двумя, тремя и т.д. значениями. Такой вид распределения представляет собой, по-видимому, универсальную квантитативно-системную характеристику полисемии естественных языков. По существу здесь, как и во многих других случаях, проявляется известный принцип концентрации и рассеяния лингвистических единиц.

Для выяснения вопроса об аналитическом выражении связи между количеством значений и долей слов с данным количеством значений нами были просмотрены опубликованные в печати данные о полисемии в английском, венгерском и русском языках.[+] Проверка показала (см. Тулдава Ю.А., 1979), что соответствие между эмпирическими и теоретическими данными достигается несколькими способами. Хорошее соответствие дает степенная функция с поправочным параметром (типа закона Ципфа-Мандель-

[+] Данные по английскому языку взяты из статьи С.М. Вишняковой (1976), где исследуется полисемия полнозначных слов на основе тезауруса Роже (всего около 30 000 слов). Материал по венгерскому языку взят из работы Ф. Паппа (1967) и охватывает все слова в толковом словаре венгерского языка (около 58 000 слов). Пример русского языка взят из исследования Ю.К. Крылова и М.Д. Якубовской (1977), причем в данном случае рассматриваются лишь глаголы, взятые из "Словаря русского языка" С.И. Ожегова (всего около 9500 глаголов).

брота), но в целях лучшей интерпретации распределения, в частности для классификации слов с разным семантическим объемом, целесообразно использовать модифицированную экспоненциальную функцию типа

$$p(m) = ae^{-b\sqrt{m}}, \qquad (3.10)$$

где $p(m)$ - доля слов с данным количеством значений, m - количество значений (семантический объем), a и b - параметры (e - основание натуральных логарифмов). Результаты вычисления ожидаемых долей слов в сравнении с наблюдаемыми величинами приводятся в табл. 3.10.

Таблица 3.10

Связь между количеством значений m и долей слов с данным количеством значений $p(m)$ по данным словарей английского, венгерского и русского языков. Вычисления по формуле (3.10).

m	Английский яз. $p(m)$ набл.	$p(m)$ ожид.	Венгерский яз. $p(m)$ набл.	$p(m)$ ожид.	Русский яз. (глаголы) $p(m)$ набл.	$p(m)$ ожид.
I	0,427	0,426	0,504	0,558	0,615	0,627
2	0,203	0,205	0,265	0,200	0,254	0,189
3	0,117	0,117	0,118	0,090	0,071	0,075
4	0,072	0,073	0,052	0,046	0,030	0,035
5	0,048	0,048	0,024	0,025	0,013	0,017
6	0,035	0,033	0,013	0,015	0,007	0,009
7	0,023	0,023	0,008	0,009	0,003	0,005
8	0,016	0,017	0,005	0,006	0,002	0,003
9	0,013	0,012	0,003	0,004	0,002	0,002
I0	0,009	0,009	0,002	0,003	0,002	0,001
II	0,0073	0,007I	0,0014	0,0017		
I2	0,0060	0,0055	0,0012	0,0012		
I3	0,0053	0,0042	0,0009	0,0008	0,001	(0,03)
I4	0,0034	0,0033	0,0007	0,0006		
I5	0,0032	0,0026	0,0007	0,0004		
>I5	0,014	(0,015)	0,002	(0,04)		
Всего	I,0	(I,0)	I,0	(I,0)	I,0	(I,0)
Параметры:		a = 2,5 b = I,77		a = 6,8 b = 2,5		a = II,4 b = 2,9

В такой интерпретации квадратный корень от количества значений ($\sqrt{m}$) становится как бы новой единицей измерения семантического объема: последовательность натуральных чисел I ($=\sqrt{1}$), 2 ($=\sqrt{4}$), 3 ($=\sqrt{9}$) и т.д. маркирует интервалы, которые можно связать с естественной группировкой слов в подклассы на основе степени полисемичности, например:

нулевая степень полисемичности - слова с I значением;

I-я степень полисемичности - слова с 2 - 4 значениями;

2-я степень полисемичности - слова с 5 - 9 значениями;

3-я степень полисемичности – слова с 10 – 16 значениями; и т.д.

Теоретическим обоснованием подобной зависимости может служить "энергетический принцип", согласно которому вероятность реализации явления убывает пропорционально экспоненте от его сложности (аналога энергии в термодинамике; см. Шрейдер Ю.А., 1967). Сложность интерпретируется в данном случае как сложность семантической структуры слова, измеряемая количеством значений (точнее, квадратным корнем от количества значений). Другой подход к классификации слов по степени полисемичности предлагает А.А. Поликарпов (1987).

Распределение по формуле (3.10) можно рассматривать как частный случай более общего распределения типа

$$p(m) = a e^{-b m^c}, \qquad (3.11)$$

где a, b и c – параметры. Наши экспериментальные данные говорят о том, что параметр $c \approx 0{,}5$ и формула (3.11) переходит в формулу (3.10), так как $m^{0,5} = \sqrt{m}$.

Следует добавить, что имеются и другие трактовки связи между семантическим объемом и количеством слов. Например, Ю.К. Крылов и М.Д. Якубовская (1977) выдвигают тезис об оптимальном распределении согласно принципу "максимального семантического содержания лексики", при котором выводится экспоненциальное распределение типа $p(m) = e^{-bm}$. Его можно также рассматривать как частный случай распределения (3.11), если $a = 1$, $c = 1$ и в "идеальном" случае $e^{-b} = 0{,}5$, т.е. вероятность того, что слово имеет m значений, убывает в геометрической прогрессии со знаменателем равным 0,5. При этом количество однозначных слов должно составить ровно половину словаря, количество слов с двумя значениями должно составлять половину количества однозначных слов и т.д.

Особый интерес представляют исследования, в которых рассматривается связь полисемии с другими структурными особенностями слов, например, с длиной слова (Fickermann I. et al., 1984; Sambor J., 1984; Köhler R., 1986), причем констатируется, что зависимость между полисемией и длиной слова подчиняется степенному закону (который названными авторами трактуется как частный случай "закона Менцерата"). Тесная корреляция обнаружена также между полисемией и словообразовательной активностью слов (см., например, Тихонов А.Н., 1983).

Связь с частотой слова. Известно, что всякое слово (в том числе многозначное слово) употребляется в тексте, как

правило, лишь в одном определенном, "актуализированном" значении. Несмотря на большое число возможных нюансов значения слова в конкретных контекстах, можно выделить определенное число однородных употреблений, которые и фиксируются в толковом словаре в виде отдельных, узуальных значений. Этим оправдывается изучение прямой связи между с е м а н т и ч е с к и м о б ъ е м о м слова по данным словаря и у п о т р е б л е н и е м этого слова в тексте. Если задаться вопросом о характере связи между количеством словарных значений и частотой употребления слова, то чисто умозрительным путем можно прийти к выводу, что эти два квантитативных признака слова должны находится в отношении прямой пропорции: чем больше значений у слова, тем чаще это слово должно употребляться в тексте (в речи). Если бы все отдельные значения слов имели одинаковый "вес" и были одинаково частыми, то отношение между количеством значений и частотностью слова можно было бы определить однозначно: слово, имеющее два значения, употреблялось бы в тексте в два раза больше, чем однозначное слово, слово с тремя значениями – в три раза больше, чем однозначное слово и т.д. Но, как известно, внутрисловное смысловое разграничение само по себе раскрывает сложную сеть значений слова, в частности, в форме частотной иерархии значений (среди значений многозначного слова обычно выделяется одно наиболее частое "основное" значение). Также могут различаться по частоте употребления актуализованные значения разных слов – в зависимости от коммуникативных нужд и требований структуры языка. Кроме того, надо учесть факт постоянного изменения и развития лексико-семантической системы языка в связи с развитием общественных отношений. Таким образом, отношение между семантическим объемом слова и его употреблением в тексте носит сложный характер, и на данном этапе исследования мы можем лишь попытаться выяснить характер этой связи в интегральном плане. При этом мы исходим из положения, что несмотря на всевозможные локальные отклонения, система в целом сохраняет относительную устойчивость в каждый данный момент своего функционирования.

В первую очередь нас интересует установление формы связи между семантическим объемом, который мы измеряем количеством значений слова в словаре, и частотой употребления слова в тексте. Практически удобно представить себе количество значений слова как функцию от средней частоты употребления (хотя, в действительности, зависимость здесь обоюдная). Для при-

мера возьмем имеющиеся данные по русскому языку (по материалам исследования А.А. Поликарпова, 1976) и по эстонскому языку (Тулдава Ю.А., 1979). Группируя слова по частотным зонам с указанием средней частоты (F) и среднего количества значений слова в данной частотной зоне (m), мы можем хорошо аппроксимировать связь между переменными с помощью степенной функции типа

$$m = \alpha F^{\gamma}, \qquad (3.12)$$

где α и γ – параметры. Для русского языка частотные зоны определяются интервалом рангов, причем средняя частота вычисляется в интервалах по 100 слов по данным ЧС русского языка (1977). Данные по эстонскому языку представлены несколько иначе: за основу берется средняя частота в интервале частот в ЧС лексем эстонской художественной прозы, например, 15 для интервала 10 – 20, и в этом интервале вычисляется среднее количество значений слов. Но разница в вычислениях не меняет существа дела, и в обоих случаях наблюдается хорошее соответствие между эмпирическими (наблюдаемыми) и теоретическими (ожидаемыми) данными в пределах опытных значений; кроме того, функция дает приемлемые результаты и при экстраполяции в сторону меньших частот. Следует отметить, что по экспериментальным данным в зоне больших частот имеет место некоторое торможение роста количества значений (такое явление отмечено также в исследовании П.Ф. Андруковича и Э.И. Крылова, 1977). Это относится к небольшому числу наиболее частотных слов, и при использовании средних величин рассматриваемая функция просто сглаживает нерегулярность в зоне больших частот.

Как известно, зависимость семантического объема (количества значений) от частоты употребления слова была впервые сформулирована Дж. Ципфом (Zipf G.K., 1949) в виде формулы $m = \sqrt{F} = F^{0,5}$ (m – число значений, F – частота употребления). В свете данных настоящего исследования можно сказать, что формула Ципфа верна в том смысле, что для выражения связи между m и F используется степенная функция типа $m = \alpha F^{\gamma}$, причем константа α считается равной 1 (в содержательной интерпретации – количество значений одноразовых слов принимается равным 1). Мы видели выше, что по экспериментальным данным константа α действительно близка 1 и в некоторых случаях можно ею пренебречь. Но в общем формула Ципфа представляет собой лишь частный случай функции $m = F^{\gamma}$,

Таблица 3.II

Зависимость между средним количеством значений слова в словаре (m) и средней частотой употребления слова в тексте (F). Вычисления по формуле (3.I2).

Русский язык				Эстонский язык			
Интервал рангов	F	m набл.	m ожид.	Интервал частот	F	m набл.	m ожид.
I–I00	4369	8,2	8,2	I0–20	I5	2,9	2,9
I0I–200	758	5,4	5,5	2I–30	25	3,2	3,4
20I–300	468	4,9	4,9	3I–40	35	4,4	3,9
30I–400	374	4,4	4,7	4I–50	45	3,7	4,2
40I–500	250	4,6	4,3	5I–60	55	4,8	4,6
50I–600	234	4,3	4,2	6I–70	65	5,I	4,9
60I–700	200	3,9	4,0	7I–80	75	5,3	5,2
70I–800	I76	4,4	3,9	8I–90	85	5,0	5,4
80I–900	I55	3,3	3,8	9I–I00	95	6,5	5,6
900–I000	I40	3,7	3,7	I0I–I50	I25	6,3	6,3
Прогноз:	I00	–	3,5	Прогноз:	I0	–	2,5
	50	–	2,9		5	–	I,9
	I0	–	2,0		2	–	I,4
	I	–	I,2		I	–	I,05
Параметры:		α = I,2				α = I,05	
		γ = 0,23				γ = 0,38	

когда $\gamma = 0{,}5$. В действительности же значение γ варьирует в зависимости от объема словаря и типа языка.

Вычисление с помощью формулы $m = \alpha F^{\gamma}$ дает линию регрессии по связи количества значений (m) и частотности слов (F). Но эта формула определяет зависимость между m и F лишь в среднем и не дает точной характеристики для каждого слова в отдельности. Это отражает одну из важных особенностей вероятностных систем, а именно то обстоятельство, что упорядоченность, структура во множестве событий лишь обобщенно, интегрально отражает состояние отдельных элементов. Кроме того, надо учесть, что линия регрессии, вычисленная на основе выборочных данных, обязательно имеет свои доверительные пределы (на выбранном уровне достоверности). Удобнее всего вычислить доверительные пределы на основе средних значений m и γ, изображая их потом в билогарифмическом масштабе (при линейной связи между логарифмами m и F). Определяя доверительные пределы по трехкратному стандартному отклонению средних, можно приблизительно считать, что истинная линия регрессии лежит в этих пределах с вероятностью не менее 99 %. Главное здесь то, что мы на основе данных о доверительных пределах линии регрессии получаем возможность решать некоторые практические задачи лексикографии на достаточно высоком уровне статистической достоверности.

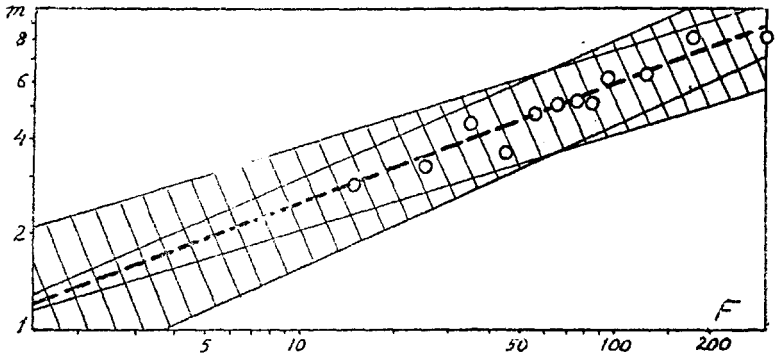


Рис. 3.4. Доверительная область для линии регрессии по связи между количеством значений (m) и частотой в тексте (F) на материале полнозначных слов эстонского языка. Билогарифмический масштаб.

Для полнозначных слов эстонского языка (при $m \approx F^{0,38}$) была построена доверительная область для линии регрессии в билогарифмическом масштабе (см. рис. 3.4; ср. табл. 3.II). (О методе вычисления доверительных пределов см., например, Пустыльник Е.И., 1968, с. 235-236).

Можно условиться о том, что доверительная область определяет зону "нормы", в которую попадают многозначные слова при допустимом колебании количества значений слов в зависимости от частоты их употребления. В тех случаях, когда отдельные слова выходят за пределы нормы (в сторону повышения или понижения количества значений), можно констатировать либо "плюсовую", либо "минусовую" полисемию. Например, в эстонском языке плюсовыми словами в отношении полисемии оказались некоторые абстрактные существительные (elu 'жизнь', meel 'чувство'), названия из животного мира (pesa 'гнездо', sarv 'рог'), из названий людей laps 'ребенок' и др. Часть этих слов часто употребляется в переносном значении, и тем самым их полисемичность повышается. Минусовыми словами в отношении полисемии можно считать отдельные высокочастотные существительные, у которых обнаруживается сравнительно узкий семантический объем, например, inimene 'человек', nägu 'лицо', päev 'день', aasta 'год', kevad 'весна'. Просмотр и качественный анализ плюсовых и минусовых слов позволяет не только выявить некоторые свойства и особенности семантического аспекта лексики, но дает также возможность проверить качество лексикографической работы.

4. СОЦИАЛЬНЫЙ И СТИЛИСТИЧЕСКИЙ АСПЕКТЫ ИССЛЕДОВАНИЯ

В главе рассматриваются вопросы социальной и функциональной дифференциации лексики и некоторые проблемы роста и развития лексики в квантитативном освещении. В разделе о стилистическом анализе основное внимание уделяется измерению лексического разнообразия ("богатства" лексики) и лексической связи текстов.

4.I. СОЦИАЛЬНАЯ ДИФФЕРЕНЦИАЦИЯ ЛЕКСИКИ

Сферы употребления лексики. Словарный состав языка дифференцируется по разным признакам, в том числе по сферам употребления, имеющим внелингвистическое, социальное основание. Словарный состав современных национальных языков распадается, как правило, на три основные группы, или разновидности: общеупотребительная (общенародная) лексика, диалектная лексика и специальная, или терминологическая лексика. Последние две группы - диалектная и специальная лексика - можно объединить под общим названием "лексика ограниченного употребления". Жаргоны и арго, относящиеся к социальным диалектам, можно рассматривать как особые подгруппы диалектной лексики (в широком смысле слова). В то же время следует рассматривать т.н. профессионализмы (полуофициальные, преимущественно разговорные разновидности терминов) как подгруппу специальной лексики.

Основные разновидности лексики существуют как реальные данности, но будучи подсистемами в составе общей системы лексики данного языка, они характеризуются всеми свойствами подсистем, т.е. они системно связаны, взаимопроникают друг в друга и т.п. Всю систему лексики можно для наглядности представить в виде пересекающихся множеств, представляющих основные компоненты словарного состава языка (рис. 4.I).

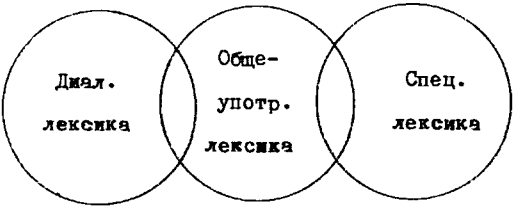


Рис. 4.I. Основные компоненты лексики современного национального языка (деление по признаку сферы употребления).

Понятие общеупотребительной лексики является в известной мере условным. К общеупотребительной лексике относятся слова, понимание и употребление которых "не ограничено какой-то местностью, профессией или родом деятельности" (Калинин А.В., 1978, с. 119). Подобная лексика составляет устойчивую основу языка, хотя по своему составу она не является однородной. В ней различаются пласты лексики устной разговорной речи (в том числе просторечия) и лексики письменной речи (в том числе "книжной" лексики). С точки зрения исторической перспективы в ней выделяются различные генетические пласты лексики, неологизмы и архаизмы. На основе статистических критериев можно выделить ядерную, или баазовую лексику (которая в основном совпадает с так называемым основным словарным фондом языка) и периферийную лексику. Во всех случаях дифференциации основным признаком общеупотребительной лексики остается то, что, по определению, ее употребление не связано с местом жительства или родом деятельности.

Нет четкой демаркационной линии между общеупотребительной лексикой и другими разновидностями лексики. Общеупотребительная лексика обогащается постоянно за счет диалектных слов, которые проникают в общее употребление в результате сближения города и деревни, через посредство художественной литературы или в результате сознательного внедрения диалектизмов языковедами. В то же время имеет место обратный процесс - влияние общеупотребительной лексики на диалекты, вследствие чего происходит сближение диалектов с общенародным языком. Общепонятными и общеупотребительными становятся со временем также многие слова социальных диалектов (жаргонные слова и арготизмы), которые употребляются в устной или письменной речи в целях экспрессии.

Нет четких границ и между общеупотребительной и специальной лексикой. В наше время наблюдается везде усиленный переход терминов в общее употребление, в результате чего во многих языках постоянно увеличивается слой общепонятных терминов. Это является одним из следствий широко распространенной "интеллектуализации" современных языков. С другой стороны, многие общеупотребительные слова приобретают наряду со своим общим значением еще специальное значение (например, 'основание' в подъязыке химии), не говоря уже о том, что большое количество терминов образуется на базе общеупотребительной лексики (например, 'корневое слово' в лингвистике).

Таким образом, приходится констатировать, что слова от-

дельных сфер употребления не имеют точно фиксированных, неподвижных разграничений, и наблюдаются постепенные, "непрерывные" переходы между лексическими пластами. Объем и границы лексических групп отдельных сфер употребления могут объективно определяться с помощью вероятностно-статистических критериев (частотность, употребительность, особенности распределения частот и др.). Известны, например, критерии разграничения общеупотребительной и специальной лексики на основе дифференцированных данных о степени употребительности слов в различных подъязыках (Андреев Н.Д., 1967) и критерии автоматического выделения терминологической лексики на основе распределения частот слов (Бектаев К.Б., 1978). Интерес представляют и эмпирические наблюдения в работах, посвященных исследованию специальных текстов, например, констатация того, что на словаре, составленном на основе специального текста объемом 100 тыс. словоупотреблений, узкоспециальные слова (относящиеся к главной тематике отрасли) концентрируются в "среднечастотной" зоне: $50 > F > 15$, где F – частота слова в тексте (Негуляев Г.А. и др.,1973).

На фоне основных групп лексики, образовавшихся в результате расчленения общей системы лексики по признаку сфер употребления, особо выделяется лексика литературного языка, или "литературная лексика", как своеобразный промежуточный слой и в то же время важнейший компонент лексических систем современных языков. Обычно считается, что литературная лексика полностью не совпадает с общеупотребительной лексикой в указанном выше значении. Она отличается от общеупотребительной лексики в целом четкой нормированностью (нормы закреплены соответствующими правилами и словарями литературного языка). В состав литературной лексики не входят просторечные слова, хотя элементы просторечия могут быть применены, например, в художественной литературе в стилистических целях. Далее, в состав литературной лексики не входят диалектные или жаргонные слова, хотя и они могут встречаться в литературных текстах как "внелитературные" вкрапления. Наконец, из литературной лексики исключаются некоторые слои специальной лексики, в частности слова профессионального просторечия. Но вопрос о включении узкоспециальной терминологии в состав литературной лексики решается по-разному. Можно, по-видимому, говорить о литературной лексике в широком и узком значении слова; в последнем случае выделяется "общеупотребительная" литературная лексика, которая включа-

ет только часть специальной лексики, фиксируемую обычно в ортографических (ортологических) или толковых словарях.

Общая и специфичная лексика. Известно, что литературный язык не является каким-то однородным целым, а разветвляется на многие жанры и стили. Например, на основе общественных функций языка (общение, сообщение, воздействие) и учитывая сферу общения (эстетическая, научная, официально-деловая и др.), в литературном языке различаются т.н. функциональные стили. Если упор делается на предметную область действительности, то говорят о "подъязыках". Иногда приходится говорить менее определенно о "жанровых группах" текстов или о текстах разных функционально-речевых сфер и т.п. Во всех этих случаях принципы формирования особых групп текстов выводятся из внеязыковых, экстралингвистических основ. Но, возникнув на внеязыковой основе, будучи тесно связанными с предметом высказывания, такие группы (и подгруппы) текстов различаются между собой и внутриязыковыми признаками, а именно особенностями отбора и употребления языковых средств. Что касается лексики этих текстов, то очевидными представляются два момента. Во-первых, основную часть словаря любого текста составляет в стилевом отношении нейтральная общеупотребительная лексика. Во-вторых, существует специфичная для каждой группы текстов лексика, которая характерна только для данной группы текстов. При сравнительном исследовании текстов, принадлежащих к разным группам, выявляется особый, ядерный слой общеупотребительной лексики, который оказывается в буквальном смысле о б щ и м для всех текстов даже при сравнительно небольших объемах выборки. Кроме того, выявляются слои лексики, общие для определенного количества текстов (групп текстов) и т.д. вплоть до специфичной лексики одной узкой группы текстов или даже одного индивидуального текста (Кожина М.Н., 1977). При таком подходе можно говорить о противопоставлении "общей" и "специфичной" лексики, но при разных уровнях общности/специфичности, выбор которых зависит от целей и задач исследования.

При квантитативном подходе во всех случаях противопоставления общей и специфичной лексики надо различать уровни словаря и текста, т.е. проводить различие между инвентарем и употреблением. Приведем пример. Частотный словарь русского языка (1977) составлен на основе текстов четырех функционально-речевых сфер (художественная проза, драматургия, газеты-журналы, научная литература), которые представлены оди-

наковым объемом выборок (по 250 тыс. словоупотреблений). Общая, совпадающая часть лексики, т.е. слова, встречающиеся во всех четырех видах текстов, составляет 6440 единиц (лексем), или 16,4 % при объеме сводного словаря 39 268 единиц. Но эти 6440 слов покрывают 82,2 % всех текстов, на основе которых составлен частотный словарь (868 577 словоупотреблений из общего числа - 1 056 382). Таким образом, выявляется большое различие между квантитативными характеристиками словаря и текста. Это различие обусловлено тем, что небольшая часть общеупотребительной лексики - в строгом смысле "общая", или "ядерная" лексика - имеет большую частотность и в количественном отношении доминирует в актуальной речи (в текстах) всех сфер общения.

Однако следует иметь в виду, что "общая" лексика ведет себя по-разному в разных сферах общения, т.е. наблюдается различие между частотными характеристиками общеупотребительных слов в разных видах текстов. Это обстоятельство позволяет (наряду с учетом количества и состава специфичной лексики) дифференцировать стили (и подстили) на основе объективных вероятностно-статистических критериев. Например, по данным упомянутого частотного словаря некоторые служебные и знаменательные слова распределяются в текстах разных видов следующим образом:

	худож. проза	драматургия	газеты, журналы	научно-технич.
а (союз)	2909	5203	1355	1252
который	673	314	1381	1600
год	246	286	1080	555
большой	297	708	571	490
сказать	1278	978	359	294

Безусловно, важное значение имеет еще тот факт, что единицы общеупотребительной лексики могут полностью не совпадать в разных функциональных стилях не только по своим количественным, но и по качественным свойствам (значение, семантический объем, экспрессивность и т.д.).

В результате различий в частотности слов выясняются и различия в распределениях, в покрываемости текста и других квантитативно-типологических характеристиках текстов, принадлежащих к разным сферам общения и тем самым к разным функциональным стилям (или подъязыкам). Если сравнивать, например, списки наиболее частотных существительных в разных функ-

Таблица 4.I

Десять наиболее частотных существительных в разных функциональных стилях эстонского, финского и русского языков

Эстонский язык				Финский язык				Русский язык	
Худож.проза		Газеты		Худож. проза		Газеты		Худож.пр.	Газеты
mees	муж(чина)	aasta	год	aika	время	vuosi	год	человек	правительство
aeg	время	valitsus	правительство	mies	муж(чина)	aika	время	рука	страна
silm	глаз	riik	государство	päivä	день	osa	часть	глаз	год
inimene	человек	rahvas	народ	käsi	рука	maa	страна; земля	дело	партия
käsi	рука	partei	партия	poika	сын; мальчик	asia	дело	жизнь	борьба
naine	жен(щин)а	poliitika	политика	ihminen	человек	ihminen	человек	голова	день
päev	день	suhe	отношение	lapsi	ребенок	työ	работа	лицо	газета
asi	дело;вещь	aeg	время	asia	дело	kysymys	вопрос	день	государство
pea	голова	osa	часть	maa	страна; земля	päivä	день	мать	сила
nägu	лицо	õigus	право	pää	голова	määrä	степень	время	вопрос

циональных стилях и языках (см. табл. 4.I), то выявляется следующее. Большое различие наблюдается в списках слов разных стилей, например, среди десяти наиболее частотных существительных в художественной прозе и в газетном тексте в эстонском языке общим является только одно слово (aeg 'время'), в русском языке - также только одно слово (день). В то же время можно констатировать близость в распределении высокочастотных существительных в одном и том же стиле, в частности в художественной прозе разных языков. Например, эстонский и финский языки имеют здесь семь общих (в данном случае генетически родственных) слов из десяти: (в переводе) мужчина, человек, рука, голова, время, день, дело. Эстонский и русский языки имеют восемь "общих", т.е. взаимопереводимых слов: человек, рука, глаз, голова, лицо, время, день, дело. Обращает на себя внимание, что среди общих высокочастотных слов художественной прозы разных языков на видном месте находятся слова, обозначающие человека (человек, мужчина, женщина) и части его тела (рука, глаз, голова, лицо). Меньше сходств в газетном тексте разных языков. Например, в эстонском и финском языках среди десяти наиболее частотных существительных-понятий общими являются только три (год, время, часть), а в эстонском и русском языках - четыре (год, правительство, партия, государство). Здесь сказывается различие в тематике, частично обусловленное социально-политическими условиями, а также тем, что рассматриваемые тексты принадлежат к разным подвидам публицистического стиля (в эстонском языке рассматриваются внешнеполитические, а в финском и русском языках - смешанные газетные тексты).

"Отмеченная" лексика. Некоторое представление о квантитативном распределении стилистических пластов слов в лексике современных языков можно получить при анализе стилистически отмеченных слов в больших толковых или других нормативных словарях. В 7-м томе "Словаря современного русского литературного языка" (1948-1965) из общего числа "позиций" - 15 530 - были отмечены пометами 3925, т.е. 25,3 % (Филин Ф.П., 1973). В "Словаре русского языка" С.И. Ожегова (1963), насчитывающем 51 533 слова, пометы употреблены 17 003 раза (Денисов П.Н., Костомаров В.Г., 1970). По приблизительным подсчетам в этом словаре должно быть около 83 000 позиций, т.е. отмеченные единицы составляют примерно 20 %. В Ортологическом словаре эстонского языка (Õigekeelsussõnaraamat, 1976), охва-

тывающем II5 000 заглавных слов, общая встречаемость помет – 42 083, или 33 % всего состава словаря. В пробной тетради толкового словаря эстонского языка (Eesti kirjakeele sõnaraamat, I969) на I6I0 заглавных слов оказалось 22II позиций, из них отмеченных 648, т.е. 29,2 % всех позиций. Таким образом, доля отмеченных единиц в рассмотренных нормативных словарях литературного языка колеблется между 20 и 33 процентами. Анализ распределения отмеченных единиц по типам стилистических характеристик выявил существенные различия в составах словарей (см. табл. 4.2). Например, в эстонских словарях среди отмеченных слов доминирует специальная лексика, в то время как в словарях русского языка преимущественно отмечена разговорная лексика. В эстонском языке нет большого различия между разговорным и общеупотребительным литературным языком и, поэтому, доля отмеченных разговорных слов ничтожна.

Таблица 4.2

Распределение стилистически отмеченных лексических единиц в словарях русского и эстонского языков (в %)

Словарь / Лексика	Сл.совр.рус. литер. яз.	Сл.рус.яз. Ожегова	Ортол.сл. эст. яз.	Толк. сл. эст. яз.
Специальная	?	I7,0	92,7	72,8
Диалектизмы	3,7	I,8	I,5	2,9
Разговорная	38,4	33,9	I,9	6,8
Просторечие	24,6	9,3		
Архаизмы	?	I3,5	I,9	5,2
Прочие	?	24,5	2,0	II,3
Всего (%)		I00,0	I00,0	I00,0
Число помет	3925 (выборка)	I7003	42083	646 (выборка)

Распределение помет по частотности в Ортологическом словаре эстонского языка приводится в табл. 4.3. Различия в частотности помет указывают на различную степень значимости отдельных групп слов, составляющих стилистически отмеченный слой в общей системе литературной лексики. Обращает на себя внимание плавное, монотонное убывание частот примененных помет, из которых большинство относится к специальной лексике. Такая регулярность в ранговом распределении частот напоминает о скрытом действии закона концентрации и рассеяния, характерных для сложных саморегулирующихся систем (ср. гл. 2.2). Если (при соблюдении принципа однородности) взять только по-

Таблица 4.3

Распределение стилистических помет в Ортологическом словаре эстонского языка. Объем словаря: 115 000 слов; число помет: 61; общая встречаемость помет в словаре: 42 083.

3433	tehn.	(техника)
2585	bot.	(ботаника)
2475	zool.	(зоология)
2437	med.	(медицина)
1991	põll.	(сельское хоз.)
1701	maj.	(экономика)
1589	sport	(спорт,физкульт.)
1465	keem.	(химия)
1442	aj.	(история)
1269	el.	(электричество)
1191	ehit.	(строительство)
1117	anat.	(анатомия)
1093	sõj.	(военное дело)
1024	lgv.	(лингвистика)
990	füüs.	(физика)
805	kõnek.	(разговорное)
799	van.	(устарелое)
744	mat.	(математика)
672	piltl.	(фигуральное)
659	biol.	(биология)
634	geol.	(геология)
625	tekst.	(текстильное дело)
623	muus.	(музыка)
611	murd.	(диалектное)
607	aiand.	(садоводство)
605	geogr.	(география)
595	pol.	(политика)
575	jur.	(юриспруденция)
546	kok.	(кулинария)
507	mer.	(морское дело)
484	farm.	(фармация)
435	kunst	(изобр.искусство)
423	mets.	(лесоводство)
413	vet.	(ветеринария)
406	etn.	(этнография)
383	kirj.	(литература)
372	trük.	(типогр. дело)
353	füsiol.	(физиология)
343	min.	(минералогия)
308	kirikl.	(церковное)
286	astr.	(астрономия)
272	folkl.	(фольклор)
262	fot.	(фотография)
227	mäend.	(горное дело)
194	teatr.	(театр)
192	meteor.	(метеорология)
182	kal.	(рыбоводство)
173	filos.	(философия)
134	ped.	(педагогика)
132	arheol.	(археология)
129	psühh.	(психология)
122	bibl.	(библиография)
92	loog.	(логика)
74	paleont.	(палеонтология)
63	müt.	(мифология)
57	vulg.	(вульгарное)
49	antr.	(антропология)
49	lastek.	(детское)
36	nalj.	(шутливое)
21	luulek.	(поэтическое)
8	halv.	(пренебрежительное)

меты специальной лексики (52 пометы из 61), то выясняется, что эмпирическое ранговое распределение их частот в целом хорошо подчиняется логарифмическому закону (см. рис. 4.2)[+], хотя наблюдается некоторое отклонение от общей тенденции в "ядре", т.е. среди наиболее частотных групп слов. Этот тип распределения встречается и в некоторых других областях лингвистики (например, в ранговом распределении частот букв в тексте).

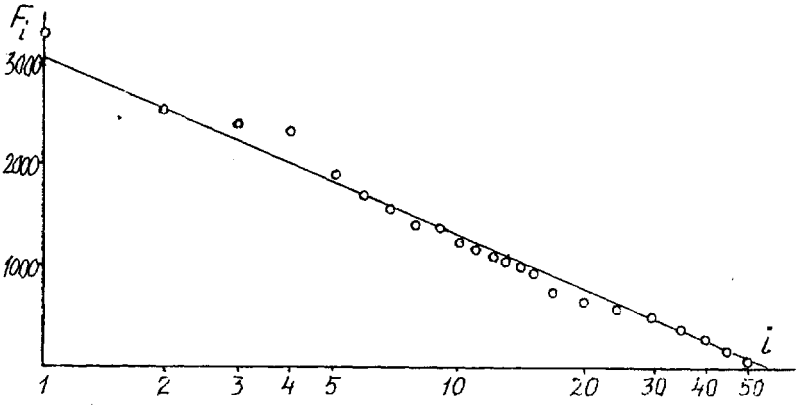


Рис. 4.2. Ранговое распределение специальной лексики в словаре. По оси абсцисс логарифмический масштаб.

Всю совокупность специальной лексики (терминов) в Ортологическом словаре можно условно сгруппировать по трем основным разделам наук (см. классификацию наук в БСЭ, т. 17, с. 330):

	Число помет	Частота	%
Технические науки	18	17 651	45,2
Естественные науки	14	12 497	32,0
Общественные науки	20	8 877	22,8
Всего	52	39 025	100,0

Распределение отдельных разделов специальной лексики в словаре безусловно отражает социальный вес и актуальность соответствующих разделов науки в наши дни.[++] Все же следует отметить, что специальная лексика общественных наук имеет

[+] Линейная зависимость между частотой (F) и логарифмом ранга ($\ln i$) по формуле $F_i = a - b \ln i$, где a и b - константы. В данном случае $a \approx 3200$ (теоретически максимальная частота), $b \approx 800$ (указывает на темп убывания частот).

[++] О статусе терминологических словарей см. Герд А.С., 1986.

большее распространение, чем это видно из приведенной таблицы, так как многие термины ежедневно встречаются в прессе, радиопередачах, телевидении и т.д., в результаты чего большое количество этих терминов перешло в разряд общеупотребительных слов и в словаре уже не отмечаются специальной пометой. Здесь сказывается принцип связи между частотностью и известностью слов.

4.2. РОСТ И РАЗВИТИЕ ЛЕКСИКИ

Модели роста словаря. Общеизвестным является тезис о том, что "в результате постоянного расширения сферы деятельности человека лексика каждого языка, особенно его терминологический словарь, несмотря на выпадение некоторого количества слов, неуклонно растет" (Пиотровский Р.Г. и др.,1977, с. 56). С математической точки зрения такому неуклонному нарастанию объема словаря соответствует в частности экспоненциальный закон роста по следующей формуле (Пиотровский Р.Г. и др., 1977, с. 57):

$$L_T = L_0 e^{kT}, \qquad (4.1)$$

где T - промежуток времени (например, столетие), L_T - объем словаря к концу периода T, L_0 - начальный объем словаря, k - коэффициент прироста, e - основание натуральных логарифмов. По экспоненциальному закону скорость роста словаря имеет "лавинообразный" характер (скорость роста словаря пропорциональна достигнутому уровню), который может быть описан следующим дифференциальным уравнением:

$$\frac{dL_T}{dT} = kL_T \quad (k>0), \qquad (4.2)$$

где k - константа. Из уравнения вытекает, что скорость роста dL_T/dT линейно зависит от достигнутого уровня L_T, а относительная скорость роста (темп прироста) $\frac{dL_T/dT}{L_T}$ остается постоянной величиной. Решая это дифференциальное уравнение, мы и получаем уравнение экспоненты (4.1).

Для проверки были взяты данные о росте лексики эстонского литературного языка на основе "представительных" словарей XVII - XX вв. (Тулдава Ю.А., 1984а). Эстонский литературный язык берет свое начало в XVI веке и проходит стадии "становления, формирования и стабилизации". Эти стадии отражаются в росте и составе представительных (наиболее полных и нормативных для своего времени) словарей (табл. 4.4).

Таблица 4.4

Рост лексики по данным словарей XVII - XX вв.

№ пп	Год	Количество слов
I.	I660	I0 000
2.	I780	I4 000
3.	I8I8	2I 000
4.	I869	50 000
5.	I893	60 000
6.	I930	I20 000
7.	I960	I05 000
8.	I976	II5 000

Анализ показывает, что экспоненциальному закону вполне соответствует рост лексики по данным рассматриваемых словарей в промежутке времени I780 - I930 (т.е. начиная со словаря № 2 и кончая со словарем № 6). По формуле (4.I) параметры L_0 = I000 и k = I,45. Соответствие эмпирических данных теоретическим хорошее (кривая I на рис. 4.3; см. также табл. 4.5; числа округлены до целых тысяч).+ По условиям экспоненциального закона роста период удвоения составляет в данном случае 48 лет++, т.е. примерно за каждое полустолетие объем словаря должен удваиваться. При таком темпе роста можно прогнозировать, что в 2000 году объем "представительного" (ортологического или толкового) словаря будет равен 330 000 словам, а в 2I00 году - полутора миллионам слов.

Очевидно, что экспоненциальный закон роста отражает какую-то реальность в том случае, если при учете объема словаря литературного языка к нему причисляются не только общеупотребительные слова, но и узкоспециальные термины, причем не учитывается выпадение слов из живого употребление (вымирание или замена слов), другими словами - если рост словарного состава рассматривается как кумулятивный процесс. Однако в жизни и в частности в лексикографической практике так обычно не бывает. Надо полагать, что безудержный рост словаря не может длиться бесконечно долго, и рано или поздно появляются задерживающие факторы, обусловленные внутренними причинами (например, насыщенностью словаря общеупотребительными словами в развитом литературном языке), так и внешними

+ За точку отсчета берется год I600 (T = 0). Далее учитывается отдаленность в столетиях от данной точки отсчета, например, при I780 (г.) T = I,8.

++ Период удвоения вычисляется как $T_y = \frac{\ln 2}{k}$. В данном случае k = I,45, следовательно, $T_y = \frac{\ln 2}{1.45}$ = 0,48, т.е. 48 лет (T_y выражает столетие).

причинами, в том числе потребностями общества и регулирующей деятельностью лексикографов.

Представляется, что рост общеупотребительной лексики и рост объема словаря литературного языка за разные промежутки времени может характеризоваться экспоненциальным законом только в отдельные периоды развития языка. В действительности процесс роста лексики начинается медленно (период становления литературного языка), затем ускоряется и принимает "лавинообразный" характер (период формирования литературного языка), но в какой-то момент процесс роста обязательно замедляется (период стабилизации). Такой схеме развития отвечает математическая модель, выражаемая т.н. логистической функцией:

$$L_T = \frac{L_\Pi}{1 + ae^{-kT}} \quad , \qquad (4.3)$$

где L_T - объем словаря к концу периода T, L_Π - теоретический предел роста словаря (асимптота), a и k - параметры функции. Графически эта модель представляется S-образной кривой, которая выражает сначала рост с возрастающей скоростью, затем скорость уменьшается и почти прекращается по мере асимптотического приближения к некоторому пределу (кривая II на рис. 4.3). Закон логистического роста действительно имеет силу в отношении роста лексики эстонского литературного языка, если исключить из рассмотрения словарь № I (1660 г.) и словарь № 6 (1930 г.; в этом словаре приводятся кумулятивно и устарелые слова). Параметры функции L_T = = 150 000, a = 280 и k = 1,8[+]. Соответствие хорошее (табл. 4.5).

Формулы (4.1) и (4.3) близки друг другу в том смысле, что при $L_\Pi \gg L_T$ (на начальных стадиях роста) кривая по формуле (4.3) практически совпадает с кривой экспоненты по фор-

---

[+] Значения параметров можно вычислить на основе линеаризации формулы (4.3). Логарифмируя, получаем: $\ln\left(\frac{L_\Pi}{L_T} - 1\right) = \ln a - kT$, т.е. линейную связь между $\ln\left(\frac{L_\Pi}{L_T} - 1\right) = Y$ и $T = X$. С помощью графика на плоскости (X, Y) устанавливается то значение L_Π, которое дает наилучшее приближение к линейной зависимости. Предварительный анализ на графике полезен тем, что отклонения от общей тенденции отдельных эмпирических точек (в данном случае № I и № 6) становятся явными, и эти точки можно при вычислении значений параметров не учитывать (отклонения получают свое объяснение). Значения параметров можно затем выявить обычным способом на графике или методом наименьших квадратов. - Установление предела по данной модели носит чисто теоретический характер, и конкретное значение L_Π не рассматривается как некий реальный предел роста словаря.

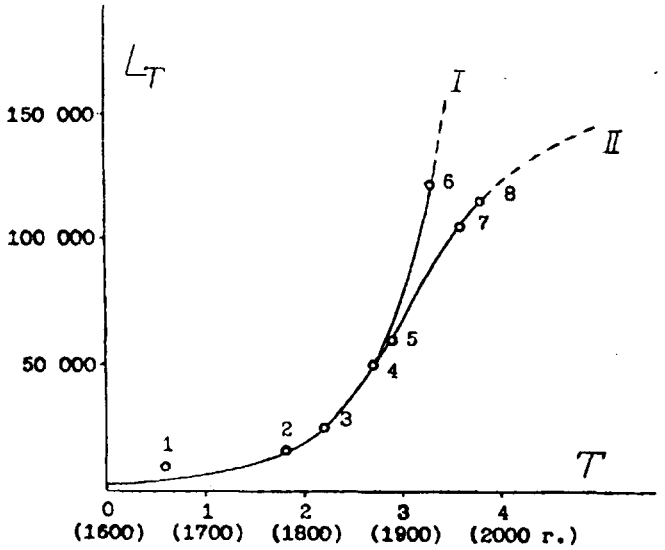


Рис. 4.3. Рост лексики эстонского литературного языка по данным словарей XVII-XX вв. Выравнивание и прогноз по экспоненциальной (I) и логистической (II) функциям. Цифры на схеме указывают на словари (табл. 4.5).

Таблица 4.5

Эмпирические и теоретические данные роста лексики на основе словарей XVII - XX вв.
(L'_T - экспоненциальное, L''_T - логистическое распр.)

№ пп.	Год	T	L_T (эмпир.)	L'_T (теор.)	L''_T (теор.)
I.	I660	0,6	I0 000	2 000	2 000
2.	I780	I,8	I4 000	I4 000	I3 000
3.	I8I8	2,2	2I 000	24 000	24 000
4.	I869	2,7	50 000	50 000	50 000
5.	I893	2,9	60 000	67 000	60 000
6.	I930	3,3	I20 000	I20 000	86 000
7.	I960	3,6	I05 000	I85 000	I05 000
8.	I976	3,8	II5 000	247 000	II5 000
-	2000	4,0	прогноз:	330 000	I24 000

муле (4.I); это хорошо видно также на графике. Таким образом, рассматриваемая кривая (по формуле (4.3)) по существу включает экспоненциальный тренд как первую стадию роста. Логистическая кривая имеет "точку перегиба", т.е. пункт перехода, где начинается непрерывное замедление скорости роста. За точкой перегиба кривая напоминает изображение логарифмической функции, отражающий "закон адаптационного торможения" (Налимов В.В., Мульченко З.М., I969). Анализ показывает, что кривая логарифмической функции по формуле

$$L_T = \alpha + \beta \ln T \qquad (4.4)$$

(в данном конкретном случае параметры $\alpha = -$ I60 000 и $\beta \approx$ 206 000) практически совпадает с кривой логистической функции по формуле (4.3) в период стабилизации развития литературного языка (примерно начиная с I900 г.).

По всей вероятности, закон логистического развития в своей общей форме (ускорение – точка перегиба – замедление) имеет всеобщее социально-лингвистическое значение и характеризует рост и развитие лексики большинства литературных языков, хотя конкретную форму этот закон принимает в зависимости от условий исторического развития данного народа – носителя языка. Можно добавить, что логистический закон роста в различных своих конкретных проявлениях (имеется ряд вариантных формул логистического роста) считается одним из основных законов развития самоорганизующихся сложных систем, если рассматривать их развитие при достаточно больших временны́х интервалах. $\int$-образной кривой характеризуются также некоторые другие диахронные лингвистические процессы ("закон Пиотровского"; см. Altmann G. et al., I983), и они находят себе в наши дни широкое применение во многих областях науки, в том числе при решении задач моделирования развития самой науки (например, Прайс Д., I966).

Изменение состава словаря. Наряду с ростом объема лексики интерес представляет и развитие лексики с точки зрения изменения ее состава с течением времени. Словарный состав языка постоянно развивается, пополняясь новыми словами и освобождаясь от старых слов, причем определенная часть остается долгое время устойчивой, неизменной. Сравнение словарей разных эпох может дать материал для выяснения закономерностей развития состава языка как в количественном, так и качественном отношении. В данном разделе рассмотрим только некоторые возможности количественного сравнения составов сло-

варей (соответствующий качественно-количественный анализ см. Тулдава Ю.А., 1984).

Для получения общего представления о масштабе изменения словарного состава за период развития эстонского национального литературного языка с конца XIX века до наших дней было проведено сравнение состава заглавных слов словарей № 5 и № 8 (см. табл. 4.4). Выборочный анализ показал, что общее число слов на букву *L* в словаре № 5 (Эстонско-немецкий словарь Ф. Видеманна, 2-е изд., 1893 г.) составляет 4700, а в словаре № 8 (Ортологический словарь 1976 г.) – 7732. Число общих для двух словарей слов оказалось 1550 (причем общими считались слова, которые совпали по форме и хотя бы по одному значению). Результаты сравнения фрагментов двух словарей можно представить в следующем виде:

Словарь 1893 г.:	Число слов в фрагменте	4700;	
	из них общих со словарем 1976 г.	1550	(т.е. 33 %);
	только в словаре 1893 г.	3150	(т.е. 67 %).
Словарь 1976 г.:	Число слов в фрагменте	7732;	
	из них общих со словарем 1893 г.	1550	(т.е. 20 %);
	только в словаре 1976 г.	6182	(т.е. 80 %).

Эти данные можно для наглядности представить в виде схемы, где фрагменты сравниваемых словарей объединяются в одно целое, выявляя таким образом общую, совпадающую часть и необщие, специфичные части словарей (рис. 4.4).

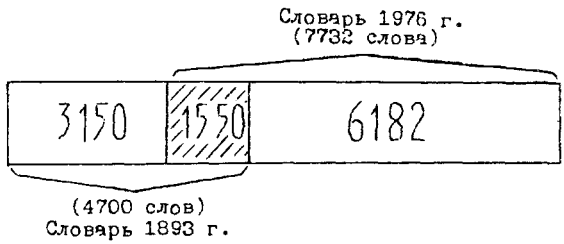


Рис. 4.4. Распределение общей и необщей лексики при сравнении (объединении) двух словарей (фрагментов). Общая часть заштрихована.

Данные квантитативного сопоставления говорят о том, что 33 %, т.е. лишь одна треть фрагмента словаря 1893 г. сохранилась до наших дней, в то время как две трети из этого фрагмента не нашли места в словаре современного ортологического

словаря. С другой стороны, в соответствующем фрагменте словаря 1976 г. сохранившиеся, т.е. перенятые из словаря 1893 г. слова, составляют около 20 %, или одну пятую часть, а все остальные являются "новыми" словами. Все это свидетельствует о глубоких качественных сдвигах в развитии литературного языка в XX веке.

Лексическую близость двух фрагментов словарей можно измерить с помощью разных индексов связи. При объединении фрагментов мы как бы накладываем их друг на друга и выявляем их общую часть и специфичные части (рис. 4.4). Исходя из этого, можно вычислить отношение общей части (C) к объему объединенного словаря ($A + B$):

$$R_I = \frac{C}{A + B - C} \quad . \qquad (4.5)$$

Индекс связи R_I может принимать значения от нуля до единицы. Очевидно, что $R_I = 0$, если $C = 0$, т.е. когда в сравниваемых словарях общих слов нет, и $R_I = 1$ при теоретически максимальной связи, а именно, когда оба словаря в точности совпадают ($A = B = C$). В нашем примере вычисление взаимной связи между словарями (фрагментами) дает следующий результат:

$$R_I = \frac{1550}{4700 + 7732 - 1550} = 0{,}142.$$

Это означает, что общая часть составляет 14,2 % объединенного словаря.

Возможны и другие варианты индексов взаимной связи. Например, можно вычислить отношение общей части (C) к среднему объему двух словарей:

$$R_{II} = \frac{C}{(A+B)/2} = \frac{2C}{A+B} \quad , \qquad (4.6)$$

что применительно к данным нашего эксперимента дает:

$$R_{II} = \frac{2 \cdot 1550}{4700 + 7732} = 0{,}249.$$

Индекс R_{II} меняется также от нуля до единицы, но он прямо не сравним с индексом R_I, так как их содержания разные. В данном случае $R_{II} = 0{,}249$ означает в содержательной интерпретации, что общие слова составляют в среднем 24,9 % объема каждого словаря (фрагмента) в отдельности.

При сильно различающихся объемах сравниваемых словарей рекомендуется взять за основу не арифметическую среднюю, как

в формуле (4.6), а геометрическую среднюю. Формула принимает тогда следующий вид:

$$R_{III} = \frac{C}{\sqrt{A \cdot B}} \quad . \qquad (4.7)$$

В нашем примере вычисление индекса по формуле (4.7)дает результат R_{III} = 0,257. Разница небольшая по сравнению с индексом R_{II} , так как объемы словарей (фрагментов) в данном случае не очень сильно отличаются друг от друга. Однако теоретически более обоснованным является применение индекса R_{III}, так как он представляет собой обобщение и включает индекс R_{II} как частный случай: если A = B, то $R_{III} = R_{II}$.

Возраст и частота слова. Во многих исследованиях последнего времени отмечается тот факт, что существует статистическая связь между "возрастом" слова, т.е. временем его возникновения в языке, и частотой употребления слова в современном языке (см., например, Арапов М.В., Херц М.М., 1974; Arapov V.M., Cherc M.M., 1983; Embleton S.M., 1986).

При квантитативно-системном подходе к исследованию связи между возрастом и частотой слова целесообразно прибегать к методу моделирования с помощью распределений. Требуется выявить такие системные свойства исследуемых объектов, которые могут служить основанием для построения распределений и их содержательной интерпретации. В данном случае оказывается наиболее целесообразным использовать методику М.В. Арапова и М.М. Херц (1974), которые в специальном исследовании рассматривают связь возраста и частоты слова для построения адекватной математической модели эволюции словаря.

По этой методике экспериментальные данные - слова в частотном словаре современного языка - объединяются в частотные зоны (группы) по 100 слов в каждой зоне. Ранжированным зонам приписываются номера, или ранги (i). В каждой зоне выявляется количество "древних" слов (т.е. слов, возникновение которых в языке датировано каким-то ранним моментом времени, например, 12-м веком) и соответствующая относительная частота, или доля.

Для примера приводим данные о распределении древних слов по частотным зонам в ЧС разных языков (табл. 4.6).+ В

+ Данные заимствованы из работы М.В. Арапова и М.М.Херц (1974) и основываются на материалах частотных словарей: по франц. яз. (автор словаря Г. Гугенейм), нем. яз. (Ф.Кединг), англ. яз. (А.Г. Дьюи), русский яз. (Э.А. Штейнфельдт), чешский яз. (А. Елинек, И. Бечка, М. Тешителова).

Таблица 4.6

Распределение слов древнего происхождения по частотным зонам (i) в шести языках

Ранг (зона) (i)	Языки					
	Эст.	Франц.	Нем.	Англ.	Рус.	Чеш.
	Датировка (год н.э.)					
	I200	I200	II00	II00	600	600
I (I-I00)	9I	9I	94	92	84	75
2 (I0I-200)	77	84	88	70	57	63
3 (20I-300)	70	84	83	53	52	50
4 (30I-400)	66	7I	73	40	5I	43
5 (40I-500)	60	73	63	47	42	37
6 (50I-600)	54	52	55	32	32	36
7 (60I-700)	46	57	55	29	42	45
8 (70I-800)	44	55	59	36	35	42
9 (80I-900)	43	52	52	3I	33	32
I0(90I-I000)	37	6I	53	3I	35	32
Всего слов древн. происх.	588	680	675	46I	463	455

Таблица 4.7

Эмпирическое и теоретическое распределение частот древних слов в трех языках; параметр a и коэффициент убывания e^{-a} по экспоненциальному закону $p(i) = e^{-ai}$; в данном случае вычислены абсолютные частоты $F(i) = 100\, p(i)$.

Ранг (i)	Эстонский		Французский		Немецкий	
(зона)	Эмп.	Теор.	Эмп.	Теор.	Эмп.	Теор.
I (I-I00)	9I	9I	9I	94	94	93
2 (I0I-200)	77	82	84	89	88	87
3 (20I-300)	70	74	84	84	83	8I
4 (30I-400)	66	67	7I	79	73	76
5 (40I-500)	60	6I	73	74	63	70
6 (50I-600)	54	55	52	70	55	66
7 (60I-700)	46	50	57	66	55	6I
8 (70I-800)	44	45	55	62	59	57
9 (80I-900)	43	4I	52	58	52	53
10 (90I-I000)	37	37	6I	55	53	50
Параметры: a	0,I		0,06		0,07	
e^{-a}	0,905		0,942		0,932	

пределах первых 1000 наиболее частотных слов наибольшее число древних слов обнаруживается во французском и немецком словарях (680 и 675, соответственно), в эстонском словаре их меньше (588), а в английском словаре значительно меньше (461). Данные этих четырех языков основываются на одинаковой датировке древних слов (1100–1200 г.н.э.), причем правила идентификации древних слов приблизительно совпадают. Для русского и чешского языков датировка древних слов относится к более раннему периоду, а именно ко времени распада праславянского единства (около 600 н.э.). Число древних слов в обоих словарях примерно одинаковое (463 и 455).

В отношении всех рассматриваемых языков можно констатировать связь возраста и частотности слов в том смысле, что доля древних слов монотонно убывает в связи с увеличением ранга частотной зоны, т.е. с уменьшением частотности в среднем. Встает вопрос о форме математической зависимости между количеством древних слов $F(i)$ или их долей $p(i)$ и рангом частотной зоны i. Эту зависимость приходится рассматривать в вероятностном плане, но формально в виде функции.

В первом приближении можно считать, что убывание количества (или доли) древних слов в связи с увеличением ранга подчиняется экспоненциальному закону, при котором средний темп убывания остается постоянным. Распределение древних слов по частотным зонам можно в таком случае представить в виде пологой экспоненты, асимптотически приближающейся к оси абсцисс (см. рис. 4.5). Аналитическое выражение такой кривой имеет вид:

$$p(i) = e^{-ai}, \qquad (4.8)$$

где $p(i)$ – относительная частота (доля) древних слов в зоне i, e – основание натуральных логарифмов, a – параметр (константа). Как известно, такую же простую зависимость постулировал М. Сводеш (1960) в своей теории глоттохронологии в отношении вероятности сохраняемости древних слов за определенные промежутки времени в истории языка.[+]

Функция (4.8) достаточно хорошо описывает динамику убы-

[+] Формула (4.8) соответствует зависимости, применяемой в археологии при радиоуглеродном датировании. Это явилось прямым стимулом для создания теории глоттохронологии. Можно еще отметить, что в формуле (4.8) компонента e^{-a} содержательно означает средний темп убывания, например, при $e^{-a} = 0{,}905$ доля древних слов составляет 0,905, или 90,5 % предыдущего уровня.

вания числа древних слов в пределах 10-12 частотных зон в эстонском, французском и немецком языках (см. табл. 4.7). Особенно хорошо эта связь проявляется в эстонском языке, где за основу был взят базовый словарь одного подъязыка (Тулдава Ю.А., 1982). На графике (рис. 4.6) видно, что соответствующее экспоненциальному закону условие линейной зависимости между $\ln p(i)$ и i хорошо выполняется.

Однако функция (4.8) не подходит для всех языков и словарей, в частности для русского, чешского и английского языков (по данным рассматриваемых словарей). М.В. Арапов и М.М. Херц (1974) предлагают обобщающую формулу типа:

$$p(i) = e^{-a\sqrt{i}} . \qquad (4.9)$$

Хотя эта формула и подходит, например, для русского языка, она все же дает слишком приблизительные оценки по материалам других рассматриваемых словарей. Приходится искать другую обобщающую формулу. Нетрудно показать, что формулы (4.8) и (4.9) являются частными случаями более общей исходной формулы с тремя параметрами c, a и b:

$$p(i) = ce^{-ai^b} . \qquad (4.10)$$

В формулах (4.8) и (4.9) параметр заранее зафиксирован: $c = 1$, т.е. он указывает на максимум вероятности древних слов. Это будет видно также при переписывании формулы (4.10) для абсолютных частот:

$$F(i) = ne^{-ai^b} , \qquad (4.11)$$

где $F(i)$ - абсолютная частота древних слов, n - объем группы (частотной зоны), т.е. максимум, или точка отсчета, на основе которого определяется вероятность $p(i) = \frac{F(i)}{n}$. В данном случае $n = 100$.

Разница между формулами (4.8) и (4.9) в том, что в первом случае жестко зафиксирован параметр $b = 1$, а во втором случае $b = 0{,}5$ (т.к. $\sqrt{i} = i^{0{,}5}$). Представляется, что целесообразнее допустить свободное варьирование параметра b, который предстает как квантитативно-лингвистический показатель, дифференцирующий языки и словари. Содержательно параметр b выражает темп убывания вероятности появления древних слов с уменьшением частотности слов. Параметр a также имеет содержательный смысл: он является показателем концентрации древних слов в начальной части частотного словаря

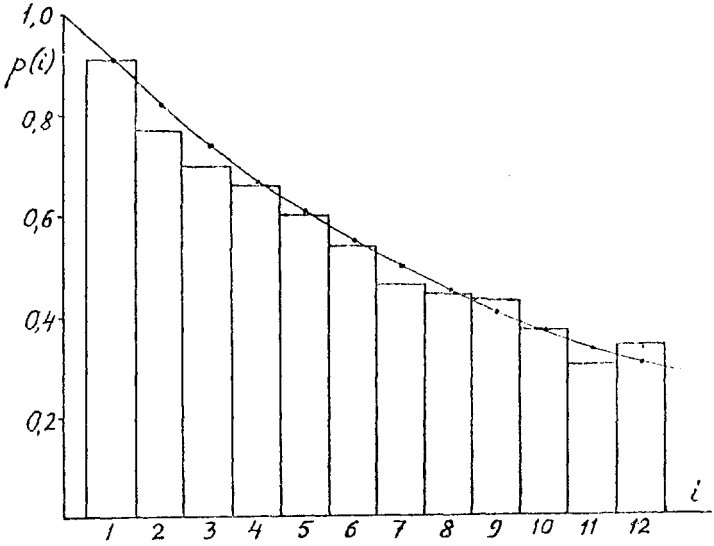


Рис. 4.5. Распределение древней лексики в частотном словаре эстонского языка. Связь между долей древних слов $p(i)$ и рангом частотной зоны (i). График функции $p(i) = e^{-0,1i}$.

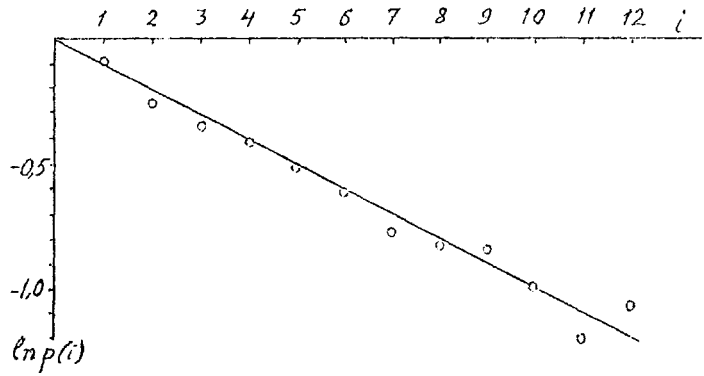


Рис. 4.6. Линейная связь между логарифмом доли $\ln p(i)$ и рангом i.

(меньшее значение a отражает бóльшую концентрацию). Уточняя значения параметров по словарю эстонского языка с помощью формулы (4.10), мы получаем: $a = 0{,}11$ и $b = 0{,}96$ (так как $b \approx 1$, то в данном случае оправдывается применение экспоненциальной функции (4.8)). Для русского языка (в пределах $i = 1 \div 25$) параметры $a = 0{,}32$ и $b = 0{,}51$ (по параметру b оправдывается применение функции (4.9)).

К вопросу аналитического описания распределения древних слов в словаре можно подойти и иначе, а именно, исходя из и н т е г р а л ь н о г о распределения слов. В таком случае надо рассматривать накопленные частоты древних слов по "вложенным" частотным зонам (в первой зоне, в первой и второй зонах вместе взятых и т.д.). При выборе соответствующей функции необходимо учесть, что рост числа древних слов имеет предел и что рост замедляется по мере приближения к пределу. Учитывая характер связи переменных по формуле (4.10), можно вывести функцию интегрального распределения:

$$p^*(i) = 1 - e^{-ai^b}. \qquad (4.12)$$

Здесь $p^*(i)$ – вероятность, соответствующая отношению $F^*(i)/F_п$, где $F^*(i)$ – накопленная частота древних слов, $F_п$ – предел числа древних слов в данной совокупности; a и b – параметры[+]. Формула (4.12) совпадает с известным в науке интегральным законом распределения Вейбулла (Weibull W., 1939; см. также Бектаев К.Б., Пиотровский Р.Г., 1973, с. 136-138).

В таблице 4.8 приводятся результаты вычислений по формуле (4.12) на материале шести языков при одинаковых условиях эксперимента (в пределах $i = 1 \div 10$ для всех языков). Можно констатировать хорошее соответствие между эмпирическими и теоретическими данными. В эстонском языке прогнозируется предел количества древних слов $F_п \approx 1000$ для данного словаря (авторской речи художественной прозы), т.е. около 7 % всего словаря (объемом 14,7 тыс. слов). Надо учесть, что $F_п$ прогнозирует число древних слов во всем словаре при условии, что темп роста остается неизменным и за пределами экспериментальных данных. Естественно, что для увеличения достовер-

[+] Параметры a и b могут быть найдены методом наименьших квадратов на основе линеаризации: $\ln\ln\frac{1}{1-p^*(i)} = \ln a + b \ln i$. Здесь $\ln a$ – начальная ордината, b – угловой коэффициент. $F_п$ определяется итеративным способом (подбирается такое значение $F_п$, при котором соответствие между экспериментальными и теоретическими данными наилучшее; на первой стадии это удобно делать на графике на основе линеаризации).

Таблица 4.8

Интегральное распределение частот древних слов в шести языках: эмпирические и теоретические данные; параметры распределения Вейбулла

Ранг (зона)	Эстонский		Французский		Немецкий		Английский		Русский		Чешский	
	Эмп.	Теор.	Эмп.	Теор.	Эмп.	Теор.	Эмп.	Теор.	Эмп.	Теор.	Эмп.	Теор.
I (I-100)	91	91	91	93	94	96	92	94	84	84	75	76
2 (I0I-200)	I68	I70	I75	I77	I82	I80	I62	I58	I4I	I42	I38	I33
3 (20I-300)	238	24I	259	255	265	257	2I5	2I2	I93	I93	I88	I84
4 (30I-400)	304	305	330	327	338	329	255	258	244	242	23I	230
5 (40I-500)	364	364	403	395	40I	395	302	300	286	282	268	273
6 (50I-600)	4I8	4I7	455	459	456	458	334	337	3I8	32I	304	3I3
7 (60I-700)	464	466	5I2	5I9	5II	5I6	363	37I	360	359	349	35I
8 (70I-800)	508	5I0	567	576	570	570	399	408	395	394	39I	387
9 (80I-900)	55I	55I	6I9	630	622	622	430	43I	428	428	423	42I
I0 (90I-I000)	588	588	680	680	675	670	46I	458	463	460	455	453
Параметры:												
F	I000		I60C		I500		900		I900		I500	
a	0,095		0,060		0,066		0,II0		0,045		0,052	
b	0,97I		0,965		0,953		0,807		0,792		0,836	

ности прогноза придется увеличить объем экспериментального материала. Но для сравнительного типологического анализа, по-видимому, достаточно охватывать экспериментом лишь первую тысячу наиболее частотных слов (см. также Арапов М.В., Херц М.М., 1974, с. 59). Показатель F_n можно в таком случае рассматривать как относительную оценку "архаичности" данного словаря. Другим параметрам распределения (4.12) можно также придать содержательный смысл. Параметр a выражает (по отношению к F_n) степень концентрации древних слов в начале словаря (большее абсолютное значение a означает относительно бóльшую концентрацию). Параметр b отражает темп роста.

Таким образом, применение функции Вейбулла в качестве модели распределения древних слов в словаре показывает, что параметры этого теоретического распределения могут в прямом или косвенном смысле служить квантитативно-лингвистическими характеристиками и стиледифференцирующими факторами при анализе лексики. С помощью функции распределения возможны также экстраполяция и интерполяция данных в пределах словаря. В общей сложности параметры распределения Вейбулла отражают как интегральные свойства, так и внутрисистемные взаимосвязи между элементами системы лексики, и вместе с тем указывают на системную связь между возрастом и частотностью слов.

4.3. ЛЕКСИКО-СТИЛИСТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

С точки зрения квантитативно-лингвистического анализа текстов можно выделить ряд актуальных проблем, связанных со стилистическим аспектом исследования лексики данного текста, в частности вопросы о "богатстве" лексики текста, о лексической связи (близости) текстов и о классификации текстов на основе квантитативных лексико-стилистических параметров.

Лексическое богатство текстов. Вопрос об объективных методах оценки лексического богатства текстов уже давно привлекает внимание исследователей, занимающихся проблемами квантитативного изучения индивидуальных и функциональных стилей (Мистрик Й., 1967; Ворончак Е., 1972; Těšitelová M., 1972; Ratkowsky D.A et al., 1980; и др.). Лексическое богатство в квантитативном смысле определяется самым общим образом как количество разных слов в тексте (т.е. объем словаря данного текста), или как отношение количества разных

словоформ (V) или лексем (L) к объему текста (N), т.е. V/N или L/N. Это отношение называют "индексом разнообразия", или "индексом TTR" (англ. type-token ratio). Чем больше значение индекса, тем больше разных слов (словоформ или лексем) употребляет пишущий или говорящий в пределах исследуемого текста.

При сравнении лексического богатства разных текстов следует иметь в виду, что прямое сравнение значений индекса возможно только при условии одинаковых объемов текстов, так как соотношение между объемом текста и объемом словаря не остается неизменным при любых значениях N (ср. гл. 2.I). В тех случаях, когда необходимо сравнивать словари текстов неодинакового объема, прибегают к помощи особых методов (см. ниже).

Важное место при оценке богатства лексики отводится также количеству редких слов в словаре исследуемого текста. Обычно вычисляется отношение числа одноразовых слов к объему соответствующего словаря или текста. Количество одноразовых слов обозначается в общем случае символом m_1 или дифференцированно – V_1 на уровне словоформ и L_1 на уровне лексем. Таким образом, индекс одноразовых слов, или "индекс исключительности" имеет вид: V_1/V или V_1/N (на уровне словоформ) и L_1/L или L_1/N (на уровне лексем). Если в каком-нибудь тексте встречается относительно большое число слов с частотой I, то это может свидетельствовать о желании автора найти образные выражения, подобрать редкие или своеобразные слова, избежать повторения слов. В таком случае большая доля слов с частотой I говорит о богатстве и разнообразии лексики данного текста. Но этот индекс исключительности сам по себе, так же как и индекс разнообразия, не является эстетическим критерием. Только дополнительный качественный анализ может установить, что́ скрывается за большим количеством одноразовых слов в тексте: хороший стиль или излишество. Иногда обилие редких слов может затруднять понимание текста или быть признаком плохого стиля, а незначительная доля одноразовых слов может быть функционально оправдана, например, при передаче спонтанной речи. При применении индекса исключительности следует также иметь в виду, что доля одноразовых слов прямо зависит от объема рассматриваемого текста.

Для примера приводим данные эксперимента сравнения лексического состава текстов разных авторов. По выборкам одинакового объема (N = I000 словоупотреблений) из авторской ре-

чи рассказов трех писателей - "Восьмое ранение" К. Симонова, "Судьба человека" М. Шолохова, "Мы - советские люди" Б. Полевого - были вычислены упомянутые выше квантитативные характеристики (табл. 4.9).+

Таблица 4.9

Квантитативные характеристики лексики по данным выборок из авторской речи трех рассказов русских советских писателей

Автор	N	L	L_1	L/N	L_1/L	L_1/N
М. Шолохов	I000	564	437	0,564	0,775	0,437
Б. Полевой	I000	524	392	0,524	0,749	0,392
К. Симонов	I000	443	297	0,443	0,67I	0,297

Из таблицы видно, что самое высокое значение индекса разнообразия определено у М. Шолохова (0,564). Это означает, что в пределах рассматриваемых выборок у М. Шолохова богаче лексика авторской речи по сравнению с другими писателями. То же самое наблюдается при сравнении индексов исключительности, причем индивидуальные различия проявляются особенно ярко на уровне текста (размах вариации 0,437 - 0,297). Значение индекса L_1/N = 0,437 указывает на то, что одноразовые слова (лексемы) покрывают у М. Шолохова 43,7 % всего текста (выборки длиной I000 словоупотреблений).

Рассмотренные выше показатели - индексы разнообразия и исключительности - можно считать основными, "стандартными" характеристиками лексического богатства. Но в квантитативной лингвистике известны еще многие другие квантитативные показатели, которые могут прямо или косвенно диагностировать "богатство" словаря в вышеуказанном смысле. Традиционно считают мерой богатства лексики параметр γ в известной формуле Ципфа $F_i = C i^{-\gamma}$ (см. гл. 2.2). Параметр γ (тангенс угла прямой в билогарифмическом масштабе) прямо указывает на "растяжение" словаря данного текста. Чем меньше значение γ, тем бо́льшим должен быть объем словаря. Однако мы знаем, что формула Ципфа обычно не описывает частотную структуру текста в полном объеме: имеются точки излома, разделяющие основную зону слов средней частотности от зоны самых частых слов и от зоны малочастотных слов. Если вычислять значения для разных частотных зон ($\gamma_1, \gamma_2, \gamma_3$; см. гл. 2.2), то оказывается, что наиболее чувствительным к лексическому богатству является γ_3 - показатель частотной структуры текста в зоне

+ Все три рассказа цитируются по книге: Русский характер. - М.: Молодая гвардия, I970.

малочастотных слов. Соответствующий эксперимент показал, что показатель γ_3 тесно коррелирует с индексами разнообразия и исключительности. Параметр γ_2 (для зоны слов средней частоты) имеет умеренную связь с индексами богатства, а квантитативная характеристика зоны самых частых слов γ_1 прямо не связана с этими индексами (подробнее см. Тулдава Ю.А.,1977).

Выше было отмечено, что при сравнении текстов неодинакового объема следует применять особые методы для определения лексического богатства текстов. Известна, например, формула П. Гиро (Guiraud P., 1954): $R = L/\sqrt{N}$, которая, однако, пригодна лишь при сравнении текстов, незначительно отличающихся друг от друга по объему. Несколько более надежные результаты можно получить при помощи коэффициента Сомерса (Somers H.H., 1966): $R = (\ln \ln L)/(\ln \ln N)$. Теоретически и практически более обоснованными являются методы, учитывающие динамику роста словаря при увеличении объема текста (например, Орлов Ю.К., 1978). Следует указать на возможность определения лексического богатства словарей разных текстов неодинакового объема с помощью формул для аналитического выражения связи между объемом словаря и объемом текста (см. гл. 2.3). Можно, например, использовать формулу (2.40), которая позволяет "выравнивать" объемы разных текстов и параметры которой подлежат содержательной интерпретации в смысле установления тенденции и формы кривой роста словаря.

Выбор подходящей формулы зависит от конкретных условий, целей и задач исследования. При сравнении небольших текстов (выборок) примерно одинакового объема можно воспользоваться простой формулой, позволяющей содержательно интерпретировать параметры формулы. Для этого подходит формула (2.36), которую можно переписать в следующей форме:

$$L = \frac{a N}{N + b}, \qquad (4.13)$$

где a и b – параметры, вычисляемые методом наименьших квадратов посредством линеаризации, причем линейная связь наблюдается между $1/L$ и $1/N$ (или N/L и N); вместо L (число лексем) может быть взято V (число словоформ).

Формула выражает особый тип дробно-линейной функции, имеющей асимптоты (см. рис. 4.7). Очевидно, физический смысл для нашего исследования имеет лишь верхний правый квадрат системы координат, так как $N \geq 0$ и $L \geq 0$. Параметр a выражает асимптоту, указывающую на предел значения L при уве-

личении N до бесконечности. В данном случае мы можем интерпретировать этот параметр как квантитативно-стилистический показатель, выражающий тенденцию роста словаря и тем самым оценивающий "потенциальное богатство" лексики при сравнении разных текстов. Второй параметр в формуле (b) определяет темп роста словаря по сравнению с увеличением объема текста (на рис. 4.7 видно, как изменение значения b непосредственно влияет на форму кривой при постоянном значении a). Целесообразно взять за основу соотношение $a/b = \varphi$, которое можно рассматривать в качестве стилистической характеристики, определяющей относительный темп роста словаря: чем больше значение φ, тем интенсивнее рост словаря на начальных стадиях текста.

Упомянутые квантитативно-стилистические показатели, характеризующие динамику роста словаря, можно рассматривать комбинированно. Если в результате множества опытов установлены средние величины параметров a и φ для определенного жанра или стиля, то эти средние величины в своих доверительных интервалах можно считать "нормой" для данного жанра или стиля. Выявляются девять комбинаций (плюс и минус обозначают соответственно "выше" и "ниже" нормы жанра, знаком равенства обозначается вхождение в доверительные границы среднего):

a	φ	a	φ	a	φ
+	+	−	+	=	+
+	−	−	−	=	−
+	=	−	=	=	=

Последняя комбинация (=/=) отражает такой лексико-стилистический тип текста, который в рамках данного жанра или стиля является наиболее типичным, так как обе квантитативных характеристики находятся в пределах нормы. Особо следует остановиться на интерпретации четырех основных типов отклонений от нормы:

$+a/+\varphi$: прогнозируется большой словарный запас, причем наблюдается высокий темп роста словаря уже на начальных стадиях накопления словаря; такое сочетание характеристик свойственно авторам, упорно и целеустремленно работающим над своей лексикой на всех этапах работы над произведением;

$+a/-\varphi$: потенциально большой словарный запас, но рост объема словаря происходит постепенно, медленно;

$-a/+\varphi$: прогнозируется небольшой общий объем словаря, но наблюдается быстрый начальный темп роста словаря, которое вскоре угасает (например, в связи с однообразием тематики);

$-a/-\varphi$: небольшой общий словарный запас и медленный рост словаря.

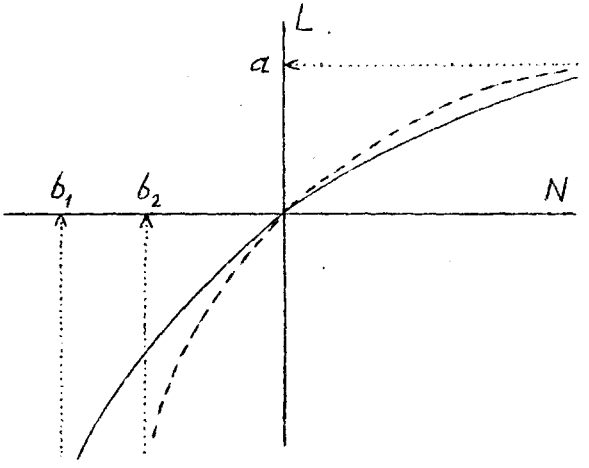


Рис. 4.7. Кривая функции $L = \frac{aN}{N+b}$ при разных значениях b и постоянном значении a .

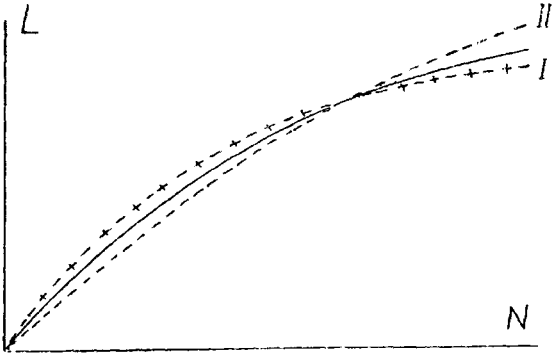


Рис. 4.8. "Норма" (сплошная линия) и варианты I, II: зависимость объема словаря (L) от объема текста (N).

Типы отклонений -/+ (I) и +/- (II) проиллюстрированы на графике (рис. 4.8).

В заключение необходимо отметить, что проблема квантитативной оценки богатства лексики, имеющая значение для стилистики и типологии текстов, не исчерпывается рассмотренными выше методами. Имеется еще ряд других аспектов проблемы, например, изучение богатства словаря в разрезе отдельных частей речи, знаменательных слов, экспрессивных слов и т.д. Во всех случаях остается в силе требование связать квантитативные данные исследования с качественным анализом.

Лексическая связь текстов. В квантитативной лингвистике существуют в настоящее время два основных подхода к измерению лексической связи (близости) двух текстов. В одном случае измеряют близость текстов по степени совпадения лексики "на уровне словаря", т.е. без учета частотности слов в тексте; при этом применяются различные индексы лексической связи (для этих целей можно, например, использовать формулы (4.5), (4.6), (4.7), примененные для сравнения составов словарей разных периодов развития языка; см. гл. 4.2). В другом случае измеряется близость текстов по степени совпадения (корреляции) частот слов в сравниваемых текстах; при этом применяются различные виды корреляционного анализа (например, Клявиня С.П., 1977; Марусенко М.А., 1981). Особой разновидностью сравнения текстов с учетом частот слов является т.н. дистрибутивно-статистический метод, при котором учитывается совместная встречаемость слов в отрывках определенной длины (Шайкевич А.Я., 1968; 1982). В настоящей работе рассматривается менее известный способ измерения лексической близости текстов с учетом частот слов - т.н. метод объединения словарей.

При сравнении лексических составов текстов можно исходить не из требования простого совпадения лексики, а из вероятностного распределения слов в том случае, если какую-нибудь исходную совокупность слов случайным образом разделили бы на две части. Из этого можно сделать вывод, что для сравнения лексики двух текстов надо провести эксперимент по объединению сравниваемых текстов (см. Muller Ch., 1968; Тулдава Ю.А., 1971; Дарчук Н.П., 1975). Предполагается, что в случае однородности лексики двух текстов (A и B) распределение подчастот в объединенном тексте подчиняется формуле $(p+q)^F$, где F - частота слова в объединенном тексте, p и q -

вероятности того, что какое-либо случайно взятое слово из объединенной совокупности относится к тексту А или к тексту В соответственно. На этой основе можно вычислить теоретические вероятности подчастот слов в текстах А и В при их объединении (см. табл. 4.IO).

Таблица 4.IO

Распределение частот в объединенном тексте по формуле $(p + q)^F$; p и q – вероятности того, что слово относится к тексту А или к тексту В соответственно; F –частота слова в объединенном тексте, f – подчастота.

F \ f	0	1	2	3	4	...	$(p + q)^F$
1	p	q					$(p + q)^1$
2	p^2	$2pq$	q^2				$(p + q)^2$
3	p^3	$3p^2q$	$3pq^2$	q^3			$(p + q)^3$
4	p^4	$4p^4q$	$6p^2q^2$	$4pq^3$	q^4		$(p + q)^4$
...							...

В самом удобном случае, когда объемы сравниваемых текстов равны ($N_A = N_B$) и, следовательно, $p = q = 1/2$, вероятности подчастот в обоих текстах принимают конкретные значения (табл. 4.II).

Таблица 4.II

Распределение частот в объединенном тексте в случае, если $p = q = 1/2$.

F \ f	0	1	2	3	4	...
1	1/2	1/2				
2	1/4	2/4	1/4			
3	1/8	3/8	3/8	1/8		
4	1/16	4/16	6/16	4/16	1/16	
...						...

Из таблицы 4.II видно, что слова с частотой 1 (в объединенном тексте) распределяются в вероятностной модели поровну между текстами А и В, т.е. 50 % отсутствует в одном тексте ($f = 0$), и столько же слов появляется I раз в другом тексте ($f = 1$). Слова, имеющие в объединенном тексте частоту 2 ($F = 2$), распределяются следующим образом: 1/4, или 25 % появляются по 2 раза в одном из текстов, например, в тексте А, и столько же отсутствуют в тексте А, т.е. появляются 2 раза только в тексте В; оставшиеся 50 % появляются по I разу в одном и по I разу в другом тексте. Таким же образом расшифровываются остальные ряды таблицы.

Особо надо отметить, что крайние величины в каждом ряду всегда соответствуют вероятности появления слов с данной частотой только в одном из текстов.

В целях иллюстрации метода мы взяли для сравнения два коротких произведения художественной литературы – рассказы "Восьмое ранение" К. Симонова и "Мы – советские люди" Б. Полевого. Из авторской речи обоих рассказов были сделаны выборки по 1000 словоупотреблений. Анализ показал, что в выборке из произведения К. Симонова – в тексте А – на 1000 словоупотреблений было 443 разных слова-лексемы, а в выборке из произведения Б. Полевого – в тексте В – 524. В объединенном тексте (при $N_A + N_B = N_{AB} = 2000$) объем словаря $L_{AB} = 841$ (см. частотные спектры индивидуальных текстов и объединенного текста в табл. 4.12).

Таблица 4.12

Частотные спектры индивидуальных текстов и объединенного текста (m_F – число слов с данной частотой)

F	Текст А $N_A = 1000$	Текст В $N_B = 1000$	Текст А + В $N_{AB} = 2000$
	m_F	m_F	m_F
1	297	392	585
2	73	68	116
3	25	26	56
4	11	12	26
5	7	6	10
6	9	3	6
7	4	2	6
8	4	–	7
9	1	2	3
⩾10	12	13	26
Всего	443	524	841

Следующим этапом эксперимента является построение теоретической и эмпирической моделей распределения подчастот и сравнение этих моделей. Теоретические подчастоты вычисляются на основе распределения частот слов, т.е. частотного спектра объединенного текста, причем исходят из теоретического распределения вероятностей подчастот в зависимости от величин p и q (в данном случае $p = q = 1/2$; см. табл. 4.11). Результаты вычислений приводятся в табл. 4.13. Распределение подчастот в эмпирической модели вычисляется прямо из сводного частотного списка слов двух текстов, например, по следующей схеме:

	A	B	AB
а	5	4	9
автомат	–	I	I
артиллерист	2	–	2
атаковать	I	I	2
большой	I	3	4

и т.д.

Эмпирическая модель для текста А представлена в табл. 4.I4 (аналогичное распределение в тексте В, если перевернуть ряды в таблице).

Сравнение теоретической и эмпирической моделей распределения подчастот в объединенном тексте выявляет некоторые интересные различия, особенно в распределении слов с малой частотой. Например, из таблицы 4.I4 видно, что слова с частотой 1 ($F = 1$) распределяются следующим образом: в тексте А – 248, а в тексте В – 337 (подчастоты I и 0 соответственно). Сравнение с соответствующими теоретическими значениями (табл. 4.I3) показывает, что в тексте А недостает 44,5 единиц, в то время как в тексте В их на такое же число больше чем в теоретической модели:

	Текст А	Текст В
Эмпирич. (Э)	248	337
Теоретич. (Т)	292,5	292,5
Э – Т	–44,5	+44,5

Для слов с частотой 2 в объединенном тексте приводим неполный список конкретных слов (табл. 4.I5). Из этого списка видно, что среди общих слов (по одному разу в каждом тексте) встречаются такие существительные, как сестра, земля, ночь, солнце, а также прилагательные медицинский, раненный. Что касается слов, встречающихся только в одном из текстов (по два раза), то из-за небольших частот невозможно точно определить, какие из них можно считать специфичными для данного текста.

В таблице 4.I6 приводится полный список слов с частотой 3 в объединенном тексте. Среди общих слов встречаются такие слова как война, фронт, врач, носилки, немецкий, что свидетельствует о близости тематики. Действительно, оба сравниваемых произведения касаются событий военного времени. Среди слов, встречающихся только в одном из текстов, можно, по-видимому, выделить два слоя – специфичные для данного текста слова, и такие, которые согласно нашей теоретической модели случайно могли оказаться только в одном из текстов. Их численное соотношение можно приблизительно рассчитать на основе эмпирических и теоретических величин, например, в тексте А

Таблица 4.I3

Распределение теоретических подчастот

F	m_F	Подчастоты									
		0	I	2	3	4	5	6	7	8	...
I	585	292,5	292,5								
2	II6	29	58	29							
3	56	7	2I	2I	7						
4	26	I,5	6,5	I0	6,5	I,5					
5	I0	0,5	I,5	3	3	I,5	0,5				
6	6	0	0,5	I,5	2	I,5	0,5	0			
7	6	0	0,5	I	I,5	I,5	I	0,5	0		
8	7	0	0	I	I,5	2	I,5	I	0	0	
...											

Таблица 4.I4

Распределение эмпирических подчастот

F	m_F	Подчастоты в тексте А									
		0	I	2	3	4	5	6	7	8	...
I	585	337	248								
2	II6	42	30	44							
3	56	I4	I2	I4	I6						
4	26	3	4	8	6	5					
5	I0	0	I	3	I	4	I				
6	6	I	I	I	0	I	0	2			
7	6	0	I	I	2	0	I	I	0		
8	7	0	0	I	0	I	2	2	0	I	
...											

Таблица 4.I5

Распределение слов с частотой $F = 2$ в объединенном тексте

	A 2	A B I I	B 2
	артиллерист гимнастерка контузия костыль маятник навещать наградить подушка пенсионный снег самолюбие самолюбивый танк шинель уйти и др.	бледный земля медицинский раненный сестра привыкнуть показаться поговорить попытаться расстаться слышать солнце ночь начало молчать и др.	барышня изящный любовь нерв омерзение палач подпольный родина сапожник смерть тюрьма утешать фашистский хатка знать и др.
Э Т	44 29	30 58	42 29
Э-Т	+I5	-28	+I3

Таблица 4.I6

Распределение слов с частотой $F = 3$ в объединенном тексте

	A 3	A B 2 I	A B I 2	B 3
	батарея выйти говорить давно если карман ни ничего остальной показывать представить полк ранить сейчас часы Чуйко	без война вместе год долго жизнь жить казаться после рассказать рука совсем товарищ ясный	взгляд врач жена идти немецкий отец работа слово такой тут фронт носилки	воля генерал дать ехать задание иной курсы летчик наш родители сам советский шеф эвакуация
Э Т	I6 7	I4 2I	I2 2I	I4 7
Э-Т	+9	-7	-9	+7
		-I6		

из общего числа слов с частотой 3 (Э = I6) теоретически "допустимы" 7 слов (Т = 7), в то время как Э - Т = 9 слов выходят за рамки случайных слов (в дальнейшем нужен качественный анализ).

Таким же образом можно проанализировать конкретные распределения слов с частотами 4, 5 и т.д. в объединенном тексте (см. Тулдава Ю.А., I97I).

Для вычисления статистической значимости различий между эмпирической и теоретической моделями и, заодно, для измерения б л и з о с т и структуры лексики двух текстов, можно использовать коэффициент близости (K) по формуле

$$K = 1 - \sqrt{\frac{\chi^2}{n+\chi^2}}, \qquad (4.I4)$$

где n - число наблюдений.

Вычисление хи-квадрата дает результат 67,I (значительно превышающий критическое значение критерия на уровне значимости 0,I %, т.е. разница между эмпирической и теоретической моделями распределений подчастот по табл. 4.I3 и 4.I4 статистически существенна). Коэффициент близости

$$K = I - \sqrt{\frac{67,I}{84I + 67,I}} = 0,728.$$

При сравнении большего числа текстов между собой можно выделить пары или группы текстов, отличающихся большей или меньшей взаимной близостью по данным распределений частот слов в тексте.

Кластер-анализ. Кластер-анализ можно определить как совокупность методов, предназначенных для разбиения некоторого множества объектов на группы, или к л а с т е р ы (англ. cluster 'группа, кучка, пучок') так, чтобы в каждой группе находились в некотором смысле наиболее близкие между собой объекты. Методы кластер-анализа относятся к группе процедур, именуемых в совокупности методами распознавания образов, а в более узком смысле методы кластер-анализа можно отнести к методам классификации многомерных наблюдений (см., например, Айвазян С.А. и др., I974). Особенностью классификации многомерных наблюдений является то, что каждый объект описывается с помощью набора (множества) зафиксированных на нем признаков, причем для построения классификации таких объектов используется данный набор признаков в их взаимосвязи.

Существует ряд разновидностей кластер-анализа, но для

них является общим наличие трех основных типов данных, используемых при проведении анализа: исходные многомерные данные, данные о близости, данные о кластерах (Крускал Дж., 1980, с. 21). Соответственно можно различать три этапа исследования: на первом, подготовительном, этапе упорядочиваются исходные данные, а на двух последующих этапах измеряется близость (сходство или различие) между классифицируемыми объектами и конструируется кластер-система, которая объединяет объекты при различных уровнях близости. Два последних этапа выполняются, как правило, с помощью автоматических процедур классификации на ЭВМ. Решением задачи кластер-анализа является разбиение, удовлетворяющее определенному критерию качества.

Необходимо подчеркнуть, что кластер-анализ, как и всякий другой метод классификации, субъективен и относителен в том смысле, что результаты анализа целиком определяются теми признаками, которые положены в его основу. Классификации, основанные на большом количестве и разнообразии признаков, будут, конечно, более эффективны для определения "естественного" порядка среди объектов и явлений (если удается использовать всю доступную информацию о признаках классифицируемых объектов). В других случаях, когда исследователя интересуют только некоторые свойства объектов, или когда кластер-анализ должен служить нуждам некоторых специальных практических приложений, можно довольствоваться небольшим числом специально отобранных признаков. В настоящей работе ставится как раз такая ограниченная задача - выявить возможности классификации текстов с помощью кластер-анализа на основе некоторых известных в практике квантитативной лингвистики формальных характеристик (лексико-)статистической структуры текста. При этом встает вопрос о сходстве результатов различных опытов, проведенных на одном и том же материале, но на основе разных наборов признаков.+

Общая задача кластеризации текстов, в том числе художественных текстов, возникает в исследованиях по изучению типологии текстов (для стилистических, педагогических и др. целей), при решении задач в области информатики, аттрибуции текстов и т.д.

+ В эксперименте использовалась ЭВМ ЕС-1060 и машинная программа кластер-анализа, разработанная Р. Эзремаа (1978).

В данной работе подвергается кластер-анализу 20 текстов - выборок по 5000 словоупотреблений из авторской речи 20 произведений современной эстонской художественной прозы (см. Тулдава Ю.А., 1981а). Считается, что выборки по 5000 словоупотреблений (каждая из которых в свою очередь разделена на 5 порций по 1000 словоупотреблений) достаточны для выявления некоторых существенных формальных показателей интересующей нас статистической организации текстов (в сравнительном плане при одинаковых объемах текстов). На материале 20 текстов было проведено три опыта на основе разных наборов измерений квантитативно-лингвистических характеристик текстов. Наборы характеристик следующие:

- покрываемость текста словоформами (опыт № 1);
- частотный спектр (опыт № 2);
- динамика роста словаря (опыт № 3).

Конкретные исходные данные приводятся в таблицах 4.17, 4.18 и 4.19.

Следует отметить, что все выше описанные наборы характеристик (т.е. данные о покрываемости текста, о частотном спектре и об объеме словаря) рассматриваются обычно как тесно связанные между собой квантитативные показатели статистической структуры текста. Вопрос состоит в том, покажет ли нам эксперимент близкие результаты в трех разных опытах классификации реальных текстов, учитывая сказанное о взаимосвязи используемых наборов характеристик.

Математической основой для классификации объектов с помощью кластер-анализа является вычисление функций на парах объектов, исходя из численных значений признаков. В результате получаются **матрицы близости** (матрицы сходства или различия) между объектами. В таких матрицах представлено множество из n подлежащих кластеризации объектов, для которых исходные (первичные) данные измерений упрощены до набора из $n(n-1)/2$ значений близости между объектами по всем парам объектов.

Задачи кластер-анализа можно решать в терминах матрицы сходства или в терминах матрицы различия. Матрицы сходства обычно конструируются на основе коэффициентов подобия или коэффициентов связи (корреляции). Матрицы различия конструируются на основе показателей "расстояния" (обзор различных мер близости см. Елисеева И.И., Рукавишников В.О., 1977, с. 31 и след.). Выбор метрики для измерения расстояния определяется природой исходных признаков и целью классификации.

Таблица 4.17

Опыт № 1. Исходные данные: покрываемость текста словоформами (%)

№ текста	Ранги словоформ (i)								
	I	I0	50	I00	500	I000	I500	2000	2500
I.	2,7	II,0	2I,4	28,3	50,6	62,6	72,6	82,6	92,6
2.	2,6	I2,0	23,3	30,4	52,4	64,2	74,2	84,I	94,I
3.	2,8	I3,8	26,7	33,7	57,5	70,7	80,7	90,7	I00,0
4.	3,5	I2,9	24,4	3I,2	53,6	66,2	75,2	85,2	95,2
5.	3,7	I3,8	25,0	3I,3	52,2	62,7	72,7	82,7	92,8
6.	2,4	I4,3	27,0	34,7	57,6	69,7	79,7	89,8	99,8
7.	3,4	I4,4	25,5	32,7	54,4	66,1	76,1	86,1	96,1
8.	3,4	I5,3	27,9	35,7	58,9	7I,I	8I,I	9I,2	I00,0
9.	3,I	I0,7	20,3	27,2	50,6	62,I	72,2	82,3	92,4
I0.	I,6	I0,I	20,2	27,I	47,2	58,5	68,6	78,6	88,6
II.	3,3	I5,4	27,0	33,9	55,5	66,8	76,7	86,6	96,5
I2.	2,6	I4,3	27,2	35,2	57,6	69,0	79,0	89,0	99,0
I3.	3,7	I4,I	27,2	34,8	56,5	68,I	78,I	88,2	98,3
I4.	3,4	I3,8	25,4	32,3	55,2	66,6	76,6	86,6	96,6
I5.	2,5	I3 I	25,4	32,I	54,5	66,I	76,2	86,4	96,6
I6.	2,9	I3,I	25,9	33,4	57,I	70,I	80,I	90,2	I00,0
I7.	3,6	I6,7	30,2	38,0	6I,I	72,5	82,5	92,5	I00,0
I8.	3,7	I3,7	27,4	35,I	58,2	70,2	80,2	90,3	I00,0
I9.	3,5	II,5	2I,4	27,6	48,9	59,9	69,7	79,4	89,2
20.	3,9	II,9	22,6	28,8	50,0	6I,4	7I,4	8I,4	9I,4

Таблица 4.18

Опыт № 2. Частотный спектр - доля словоформ (%) с данной частотой

№ текста	Частота словоформы (F)											
	I	2	3	4	5	6	7	8	9	I0	II-20	>20
I.	79,I2	I0,32	4,II	I,95	0,98	0,80	0,66	0,2I	0,38	0,2I	0,77	0,49
2.	78,70	I0,65	4,22	I,93	0,96	0,64	0,43	0,39	0,25	0,22	I,04	0,57
3.	73,29	I3,60	4,22	3,2I	I,38	0,89	0,8I	0,37	0,6I	0,04	0,65	0,93
4.	78,74	I0,45	3,94	2,0I	I,24	0,9I	0,37	0,33	0,29	0,I8	0,99	0,55
5.	8I,85	9,40	2,83	2,I7	0,94	0,66	0,28	0,25	0,28	0,II	0,9I	0,52
6.	76,00	I2,I6	4,38	I,83	I,04	0,80	0,68	0,44	0,32	0,36	I,03	0,76
7.	78,28	I0,97	4,30	I,74	I,08	0,63	0,30	0,59	0,4I	0,22	0,96	0,52
8.	75,20	I2,46	4,39	2,05	I,23	0,94	0,53	0,53	0,33	0,33	I,3I	0,70
9.	80,I8	I0,05	3,72	I,32	I,II	0,80	0,73	0,52	0,25	0,I7	0,63	0,52
I0.	8I,74	I0,08	3,36	I,47	0,65	0,52	0,39	0,23	0,I3	0,29	0,72	0,42
II.	78,6I	I0,85	3,96	I,9I	0,82	0,7I	0,75	0,4I	0,22	0,I9	0,97	0,60
I2.	77,70	II,25	3,96	I,53	I,I0	0,74	0,47	0,59	0,59	0,I9	I,25	0,63
I3.	77,80	II,29	4,60	I,24	I,08	0,74	0,50	0,23	0,27	0,43	I,I2	0,70
I4.	78,57	9,89	4,I9	2,2I	I,3I	0,94	0,49	0,34	0,30	0,22	0,94	0,60
I5.	78,64	I0,49	3,90	2,I7	I,27	0,90	0,38	0,22	0,I9	0,I9	0,94	0,7I
I6.	74,07	I2,8I	5,I8	2,09	I,25	0,96	0,80	0,52	0,32	0,I2	I,08	0,80
I7.	75,73	II,38	5,I0	I,98	I,I4	0,84	0,59	0,42	0,46	0,34	I,22	0,80
I8.	76,I2	II,44	4,43	2,42	I,4I	0,64	0,56	0,32	0,48	0,24	I,I3	0,8I
I9.	8I,56	9,0I	3,54	I,90	I,08	0,62	0,59	0,46	0,I6	0,07	0,72	0,39
20.	80,60	I0,33	3,I4	I,77	I,I9	0,68	0,34	0,38	0,I4	0,24	0,68	0,5I

Таблица 4.19

Опыт № 3. Динамика роста словаря (число разных словоформ при различных объемах текста)

№ текста	Автор	Объем текста (N)				
		1000	2000	3000	4000	5000
I.	Э. Бакман	73I	I383	I865	2404	2869
2.	В. Гросс	677	I3I5	I859	2358	2791
3.	А. Хант	649	III6	I597	2034	2463
4.	Х. Кийк	7I0	I35I	I828	2315	2738
5.	Я. Кросс	723	I3I5	I9I4	2362	286I
6.	П. Куусберг	645	II66	I674	2075	2508
7.	Дж. Промет	674	I2I2	I720	2207	2694
8.	В. Саар	633	II28	I700	2045	2439
9.	Х. Сергo	734	I325	I885	24I6	2876
I0.	Р. Сирге	764	I397	2017	2572	3067
II.	М. Траат	689	I235	I722	2225	2656
I2.	Э. Вётемаа	680	I208	I734	2162	2552
I3.	А. Каал	65I	I204	I690	2II9	2586
I4.	Т. Каллас	663	I226	I733	2I79	2668
I5.	Ю. Пээгель	690	I224	I700	2223	2669
I6.	Ю. Туулик	624	II35	I560	2006	2491
I7.	А. Халтон	586	I036	I488	I955	2373
I8.	М. Унт	658	II76	I678	2I39	2483
I9.	Э. Кийт/Я. Кросс	740	I359	I923	2473	2969
20.	Ю. Сяуля	732	I36I	I9I7	2473	2929

В данном исследовании мерой близости было выбрано обычное евклидово расстояние, исходя из следующих содержательных соображений: при данных наборах признаков и при равных объемах текстов все значения признаков (т.е. отдельные компоненты вектора) можно считать равноправными, и численные различия между отдельными значениями признаков сравниваемых текстов можно считать существенными для определения расстояния между текстами. Однако для того, чтобы избежать слишком большого веса больших численных значений отдельных признаков по сравнению с малыми значениями, необходимо выравнивать диапазоны изменения значений признаков с помощью нормализации исходных данных (обычным способом, т.е. вычитанием среднего и делением на стандартное отклонение, так что дисперсия оказывается равной единице, см. Дюран Б., Оделл П., 1977, с.40). Евклидово расстояние (d) определяется формулой:

$$d(X_s, X_t) = \left[\sum_{j=1}^{k} (x_{js} - x_{jt})^2\right]^{0,5} \qquad (4.15)$$

где x_{js} и x_{jt} – нормированные значения признаков, k – число измерений. Значение $d(X_s, X_t)$ для заданных векторов X_s и X_t считается эквивалентным расстоянию между самими объектами (текстами) T_s и T_t соответственно выбранному набору признаков. Предполагается, что близость между текстами свидетельствует о б л и з о с т и с т и л е й авторов в отношении некоторых скрытых для прямого наблюдения индивидуальных особенностей, выражающихся в устойчивых квантитативных (лингвостатистических) характеристиках текста.

При конструировании кластер-системы, или кластеризации, исходят из данных о близости между объектами, то есть, образно говоря, в алгоритмах кластеризации матрицу близости берут в качестве входа, а разбиение на кластеры является выходом. Методы кластеризации можно разделить на иерархические и неиерархические (обзор наиболее известных разновидностей кластер-анализа см. Айвазян С.А. и др., 1974, с. 99 и след.). Иерархические процедуры кластеризации бывают двух типов – агломеративные и дивизимные (разделительные). Принцип работы агломеративных алгоритмов состоит в последовательном объединении в кластер сначала самых близких, а затем и все более отдаленных друг от друга объектов. В разделительных иерархических процедурах, наоборот, множество объектов последовательно разбивается на группы. В данном исследовании была выбрана разновидность агломеративного иерархического метода кластеризации. Для практического решения вопроса в ЭВМ был

использован метод B_k, представляющий собой усовершенствованный вариант т.н. Кэмбриджского алгоритма (см. Зэремаа Р., 1978). При использовании метода B_k можно обобщенно говорить о k- кластеризации, характеризуя параметром k допустимую покриваемость кластеров до k элементов. Если имеется n объектов, подлежащих кластеризации, то параметр k может принимать целочисленные значения из отрезка $[1, n-2]$. Отметим, что при $k = 1$, т.е. при 1-кластеризации (совпадающей в данном случае с методом "одной связи", или "ближайшего соседа") получаются непересекающиеся кластеры, и их можно представить в виде дендрограммы (диаграммы-дерева). При $k>1$ это уже невозможно. В данной работе используется 1-кластеризация. Важным фактором при проведении анализа является уровень классификации, обозначаемый символом h (подробнее см. Зэремаа Р., 1978).

Как уже было сказано, при использовании агломеративного иерархического метода разбиение объектов на кластеры совершается ступенчато. Процесс кластеризации начинается с того, что (на 1-м шагу) два наиболее близко расположенных объекта (в первом опыте тексты № 6 и № 12) объединяются и рассматриваются как один кластер. Это приводит к тому, что число объектов уменьшается и становится равным $n-1$, причем один кластер будет содержать два объекта, а $n-2$ остальных по одному. Процесс можно повторять до тех пор, пока все объекты не сгруппируются в один большой кластер. Результаты такого процесса обычно изображаются графически в виде диаграммы-дерева, или дендрограммы, и с помощью отдельных таблиц с результатами кластеризации на каждом шагу (дендрограммы и таблицы выдаются в готовом виде в ЭВМ). Дендрограмма дает возможность наглядной интерпретации всего хода последовательной кластеризации (по данным наших опытов см. рис. 4.9 – 4.11). Весь процесс кластеризации в данных опытах заканчивается на 19-м шагу (при $n = 20$), где все объекты (тексты) объединяются в один кластер.

Напомним, что в трех опытах, проведенных на одном и том же материале 20 текстов, были использованы такие наборы признаков (покриваемость текста, частотный спектр, динамика роста словаря), которые обычно считаются взаимосвязанными и близкими показателями статистической структуры текста. Следовательно, можно было ожидать и близких результатов кластер-анализа по данным трех опытов. Если сравнивать соответствующие дендрограммы последовательной кластеризации (см. рис.

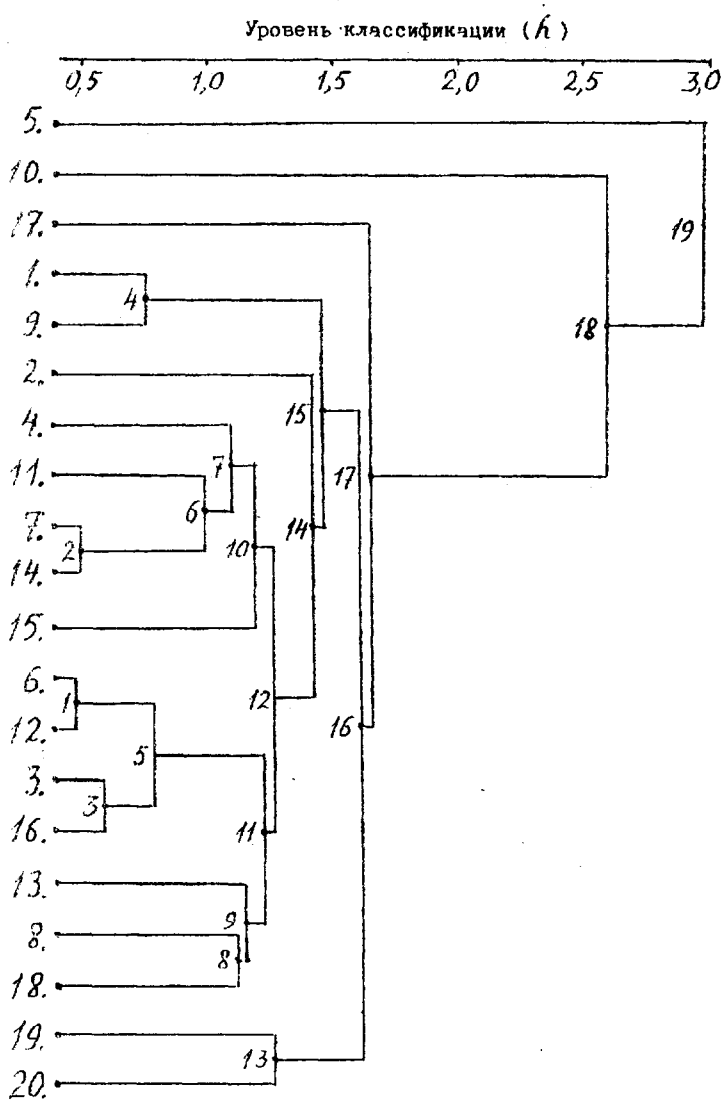


Рис. 4.9. Дендрограмма последовательной кластеризации 20 текстов на основе сравнения показателей покрываемости текста словоформами (опыт № I). Цифры слева – номера текстов. Цифры в схеме – номера шагов объединения текстов в группы.

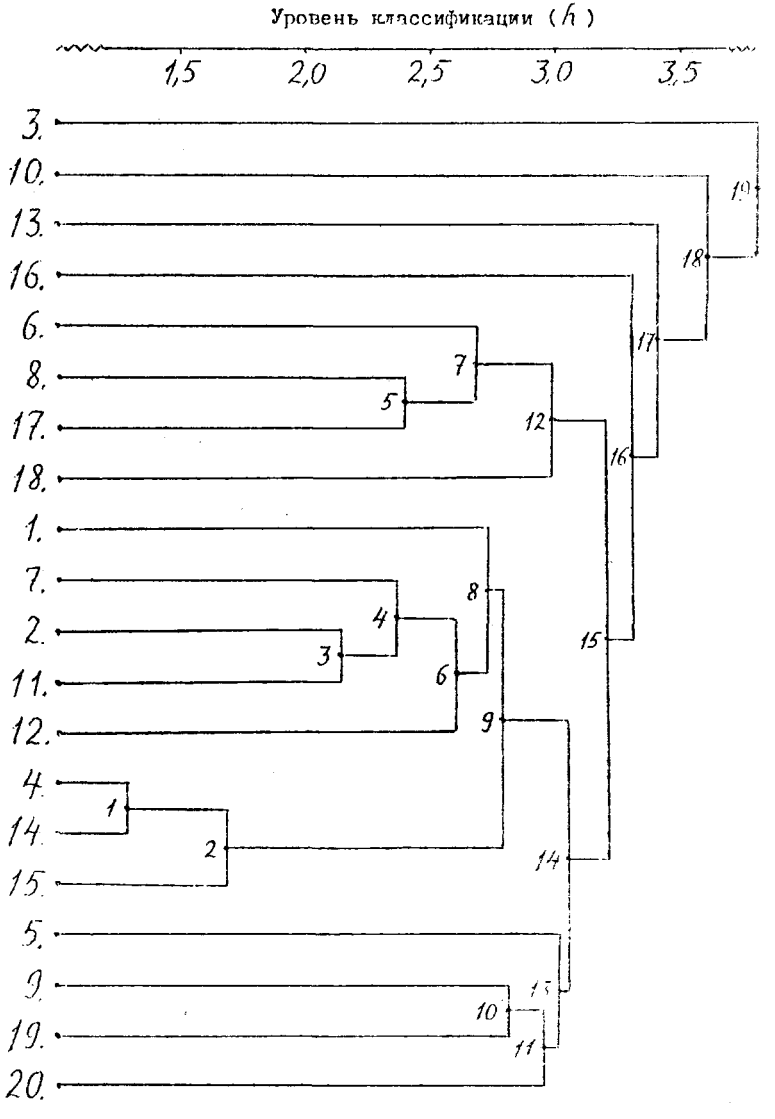


Рис. 4.10. Дендрограмма последовательной кластеризации 20 текстов на основе сравнения частотных спектров на уровне словаря (опыт № 2).

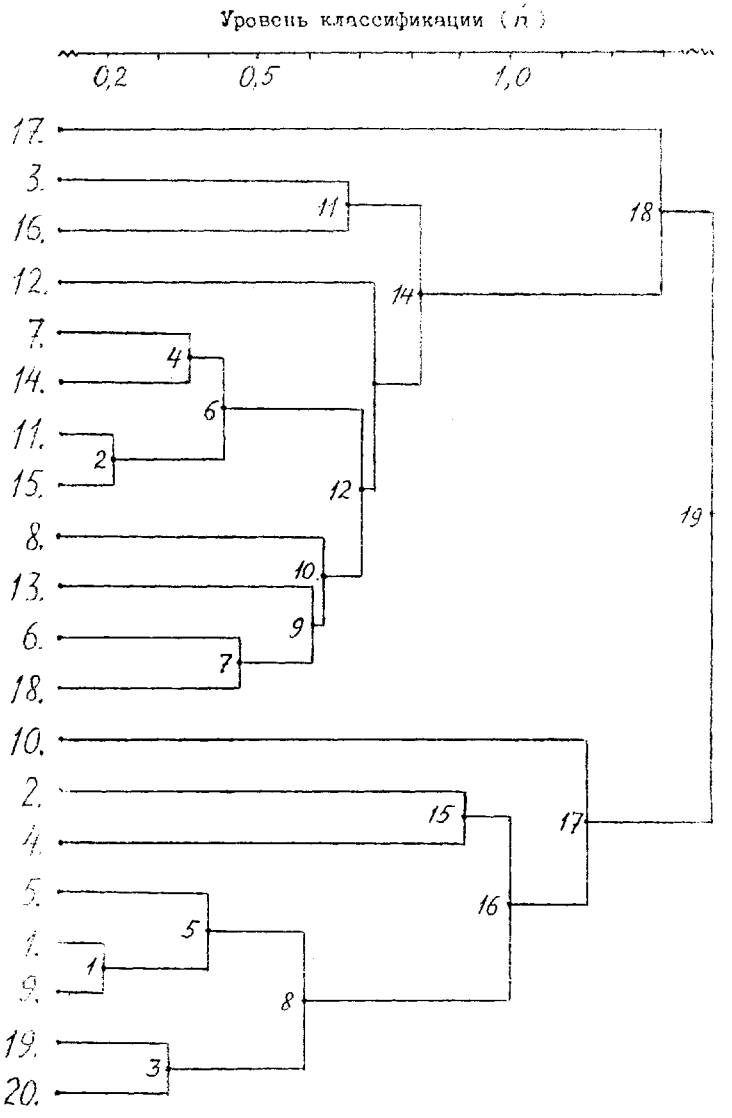


Рис. 4.II. Дендрограмма последовательной кластеризации 20 текстов на основе сравнения динамики роста словаря в связи с увеличением текста (опыт № 3).

4.9 - 4.II), то на первый взгляд удается обнаружить сходство только в отдельных точках: например, в первом и третьем опытах тексты I. и 9., а также тексты 7. и I4. объединяются в один кластер на ранней стадии кластеризации (на 4-м и I-м шагах и на 2-м и 4-м шагах соответственно). Но в общем приходится констатировать, что структуры дендрограмм мало похожи друг на друга. Поэтому постараемся сравнить такие уровни (стадии) кластеризации, которые на основе определенных критериев могут считаться "оптимальными". В данном случае желаемые (оптимальные) уровни разбиения определяются эмпирически на основе сравнения данных и с учетом оценок единичной и средней "стабильности" кластеров на отдельных уровнях разбиения (такие оценки выдаются автоматически по примененной программе ЭВМ). На этом основании можно представить следующие наборы кластеров по данным трех опытов:

	Опыт № I	Опыт № 2	Опыт № 3
"А":	(4.7.II.I4.I5.)	(I.2.4.7.II.I2.I4.I5.)	(7.II.I4.I5.)
"Б":	(8.I3.I8.)	(6.8.I7.I8.)	(6.8.I3.I8.)
"В":	(I.9.)	(5.9.I9.20.)	(I.5.9.I9.20.)
"Г":	(3.6.I2.I6.)	-	-

Изолированные тексты (I-элементные кластеры):

(2.)(5.)(I0.)	(3.)(I0.)(I3.)(I6.)	(2.)(3.)(4.)
(I7.)(I9.)(20.)		(I0.)(I2.)(I6.)
		(I7.)

При сравнительном анализе обнаруживается, что существуют отдельные общие моменты как в образовании, так и в составах кластеров.

Кластер, условно названный кластером "А", выступает в близких вариантах во всех трех опытах. Устойчивым ядром кластера "А" являются тексты 7., II., I4., I5., к ним примыкает текст 4., который встречается в кластере "А" в первом и втором опытах.

В кластере "Б" общими для всех опытов являются тексты 8. и I8., к ним примыкают тексты 6. и I3.

В кластере "В" по данным второго и третьего опытов встречаются тексты 5., 9., I9., 20.; в первом опыте тексты 5., I9. и 20. остаются на выбранном уровне классификации изолированными, т.е. составляют I-элементные кластеры, но можно констатировать, что тексты I9. и 20. через несколько шагов объединяются в один кластер (см. рис. 4.9). В то же время текст 5. остается изолированным до последнего шага кластеризации.

Кластер "Г" устанавливается только в первом опыте, в него входят тексты 3., 6., I2., I6. Но этот кластер образуется на раннем уровне, на 5-м шагу, и остается неизменным до II-го шага, что свидетельствует о большой устойчивости кластера.

Кроме многоэлементных кластеров представляют интерес и I-элементные кластеры (на выбранном уровне классификации). Во всех трех опытах неизменно изолированным остается текст I0. Тенденцию к изоляции обнаруживают также тексты 2., 3. и I7., которые составляют I-элементные кластеры в двух случаях из трех.

Итак, с помощью параллельных опытов кластеризации 20 текстов на основе разных наборов формальных характеристик статистической структуры текста удалось выделить некоторые достаточно устойчивые непересекающиеся кластеры, которые в определенной степени представляют характерные для данного языка (или подъязыка) квантитативно-лингвистические т и п ы текстов. Однако примененный метод не позволяет охватывать типизацией все тексты: в среднем 30 % текстов не попадают в устойчивые многоэлементные или I-элементные кластеры. Отчасти это может быть объяснено "эффектом сцепления" кластеров, для преодоления которого приходится останавливать процесс кластеризации на довольно раннем уровне. Но основной причиной неполного разбиения текстов на непересекающиеся кластеры следует все же считать то, что в принципе "большинство реальных классов размыты по своей природе в том смысле, что переход от принадлежности к непринадлежности для этих классов скорее постепенен, чем скачкообразен" (Заде Л.А., I980, с. 208). Таким образом, будет целесообразно основывать алгоритмы кластер-анализа на представлении о кластере (классе, типе) как о размытом, нечетком множестве (об опыте проведения кластер-анализа с частично покрываемыми классами см. Лийв Х., Тулдава Ю., I987).

При сравнительном анализе результатов трех параллельных опытов можно было констатировать значительное различие в иерархических структурах кластер-систем (см. соответствующие дендрограммы). Это различие обусловлено в большой степени тем, что наборы признаков, считающиеся близкими и тесно взаимосвязанными, в действительности не обнаруживают такого соответствия, которое необходимо для более точных расчетов. В реальных текстах нет жесткой связи между разными характеристиками статистической структуры текста. Из этого следует, что

кластер-анализ на основе одного какого-нибудь набора признаков, характеризующих статистическую организацию текста, не предопределяет результаты анализа на основе другого набора аналогичных (родственных, близких) признаков, хотя некоторое (не предсказуемое) сходство между результатами анализов имеется. При этом только сходные или совпадающие результаты параллельных опытов можно считать достаточно достоверными.

З А К Л Ю Ч Е Н И Е

На основании проведенного анализа был разрешен ряд теоретических и практических задач, касающихся разработки квантитативно-системного подхода к комплексному изучению лексики языка. К в а н т и т а т и в н о - с и с т е м н ы й п о д х о д представлен как обобщение идей и методов современной квантитативной лингвистики применительно к исследованию лексики, рассматриваемой в качестве системного объекта, в сочетании с применением расширенного арсенала методов "количественной" математики (теории вероятностей, математической статистики, теории информации, математического анализа) и имеющий целью выявление и осмысление системных свойств, т.е. таких свойств, которые могли бы служить основанием для обнаружения в лексике специфических квантитативных (вероятностно-статистических) закономерностей.

В работе был разработан и практически применен единый исследовательский аппарат, соединяющий два основных метода описания и объяснения эмпирического материала – группировку (классификацию, кластеризацию) и моделирование с помощью распределений. Практически был осуществлен поиск вероятностно-статистических зависимостей организации системы лексики и ее подсистем в виде функций (регрессионных уравнений) и показаны взаимосвязь и взаимозависимость в системе лексики в виде комплексных, многомерных распределений. Модели распределения рассматривались не просто как абстракции, а как системы, определенным образом содержательно интерпретированные. В результате анализа обширного эмпирического материала разных языков был установлен ряд новых квантитативно-лингвистических закономерностей или уточнены и обобщены ранее известные закономерности (по статистической организации словаря и текста, по фонетическим, грамматическим, семантическим и стилистическим аспектам квантитативного исследования лексики). В частности было установлено, что основные квантитативные закономерности организации словаря и текста описываются известными функциями типа закона Ципфа (степенной функции) в ранговой и спектральной формах; в более простых случаях – экспоненциальной или обратной к ней логарифмической функцией

(например, распределение лексико-фонетических групп), а также обобщенной формулой Вейбулла, которая включает экспоненциальную функцию как частный случай (например, распределение древних слов в словаре); в некоторых особых случаях - S-образной логистической функцией (рост лексики в диахронном плане) или более сложными, комбинированными функциями (например, при прогнозе роста словаря в связи с увеличением объема текста).

В заключение важно подчеркнуть, что квантитативно-системный подход к изучению и объяснению языковых явлений - еще не завершенная и складывающаяся область "на стыке наук", и по мере продвижения исследований совокупность теоретических понятий, приемов и методов работы безусловно будет возрастать и уточняться, а проблематика углубляться и обновляться.

ЛИТЕРАТУРА

Айвазян С.А., Бежаева З.И., Староверов О.В. Классификация многомерных наблюдений. – М.: Статистика, 1974. – 240 с.

Алексеев П.М. Частотные словари и приемы их составления. – В кн.: Статистика речи. – Л.: Наука, 1968, с. 61–63.

Алексеев П.М. Статистическая лексикография. – Л.: ЛГПИ, 1975. – 120 с.

Алексеев П.М. Квантитативная типология текста. АДД. Л., 1977.

Алексеев П.М. О нелинейных формулировках закона Ципфа. – Вопросы кибернетики. Вып. 41. М.; Л., 1978, с. 53–65.

Алексеев П.М. К основам статистической лексикографии. – В кн.: Проблема слова и словосочетания. Л.: ЛГПИ, 1980, с. 93–105.

Алексеев П.М. О квантитативной типологии текста. – Учен. зап. ТГУ, вып. 591. Тарту, 1981, с. 3–13.

Алексеев П.М. Методика квантитативной типологии текста. – Л.: ЛГПИ, 1983. – 76 с.

Алексеев П.М. Лингвистические распределения (Элементы количественного анализа текста). – Л.: ЛГПИ, 1985. – 56 с.

Алексеев П.М. Распределение лексических единиц по длине в тексте и словаре. – Учен. зап. ТГУ, вып. 745. Тарту, 1986, с. 3–28.

Алимов Ю.И. Альтернатива методу математической статистики. – М.: Знание, 1980. – 64 с.

Андреев Н.Д. Статистико-комбинаторные методы в теоретическом и прикладном языковедении. – Л.: Наука, 1967. – 403 с.

Андрукович П.Ф., Королев Э.И. О статистических и лексико-грамматических свойствах слов. – НТИ, сер. 2, 1977, № 4, с. 1–9.

Андриенко В.М. К вопросу об использовании коэффициента стабильности в качестве меры употребительности. – В кн.: Исследования в области вычислительной лингвистики и лингвостатистики. – М.: МГУ, 1978, с. 3–40.

Анохин П.К. Теория функциональной системы как предпосылки к построению физиологической кибернетики. – В кн.: Биологические аспекты кибернетики. М., 1962, с. 74–91.

Арапов М.В. Продуктивность в естественном языке и ее измерение. – Вопр. информац. теории и практики. Сб. 23. М.,1975, с. 117–138.

Арапов М.В. Системный анализ лексической структуры текстов. – Системные исследования. Ежегодник 1980. М., 1981, с. 372–403.

Арапов М.В., Ефимова Е.Н. Понятие лексической структуры текста. – НТИ, сер. 2, 1975, № 6, с. 3–7.

Арапов М.В., Ефимова Е.Н., Шрейдер Ю.А. О смысле ранговых распределений. – НТИ, сер. 2, 1975, № 1, с. 9–20; № 2, с. 9–20.

Арапов М.В., Тер-Гаспарян Л.И., Херц М.М. Сравнение частотных словарей. – НТИ, сер. 2, 1978, № 4, с. 20–29.

Арапов М.В., Херц М.М. Математические методы в исторической лингвистике. – М.: Наука, 1974. – 168 с.

Арапов М.В., Шрейдер Ю.А. Классификация и ранговые распределения. – НТИ, сер. 2, 1977, № 11, с. 15–21.

Арапов М.В., Шрейдер Ю.А. Закон Ципфа и принцип диссимметрии системы. – Семиотика и информатика. Вып. 10. М., 1978, с. 74–95.

Ахабаев А. Статистический анализ лексико-морфологической структуры казахской публицистики. АКД. Алма-Ата, 1971.

Ахманова О.С. Словарь лингвистических терминов. – М.: Советская энциклопедия, 1966. – 608 с.

Бартков Б.И. Количественная морфемография (дериватография) английского, немецкого, французского и русского языков. – В кн.: Основосложение и полуаффиксация в научном стиле и литературной норме. Владивосток, 1982, с. 27-55.

Бартков Б.И. Количественные методы в дериватологии. – В кн.: Исследования деривационной подсистемы количественным методом. Владивосток, 1983, с. 3-40.

Бектаев К.Б. Статистико-информационная типология тюркского текста. – Алма-Ата: Изд-во Наука Каз. ССР, 1978. – 184 с.

Бектаев К.Б., Лукьяненков К.Ф. О законах распределения единиц письменной речи. – В кн.: Статистика речи и автоматический анализ текста. – Л.: Наука, 1971, с. 47-112.

Бектаев К.Б., Пиотровский Р.Г. Математические методы в языкознании. Ч. I. Алма-Ата, 1973. Ч. 2. Алма-Ата, 1974.

Белоногов Г.Г. О некоторых статистических закономерностях в русской письменной речи. – ВЯ, 1962, № I, с. 100-101.

Белоногов Г.Г., Новоселов А.П. Некоторые количественные закономерности в автоматизированных информационных системах. – Автоматическая переработка текста методами прикладной лингвистики. Материалы Всесоюзной конференции. Кишинев, 1971, с. 219-220.

Белоногов Г.Г., Фролов Г.Д. Эмпирические данные о распределении букв в русской письменной речи. – Проблемы кибернетики, вып. 9. М., 1963, с. 287-305.

Белоногов Г.Г., Загика Е.А. и др. Автоматизация лингвистической обработки словарей. – НТИ, сер. 2, 1983, № II, с.20-24.

Белоногов Г.Г., Кузнецов Б.А., Новоселов А.П. Автоматизированная обработка научно-технической информации. – М.: ВИНИТИ, 1984.

Белоногов Г.Г., Самоделкина С.А. и др. Словообразовательные классы русских слов. – НТИ, сер. 2, 1985, № I2, с. 22-24.

Беляева Т.М., Васильева Н.М. Словообразовательное гнездо в словаре и его функциональная нагрузка в речи. – В кн.: Дериватология и дериватография литературной нормы и научного стиля. Владивосток, 1984, с. 28-35.

Бенвенист Э. Общая лингвистика. / Перев. с франц. М., 1974.

Бернштейн Н.А. Очерки по физиологии движений и физиологии активности. – М.: Медицина, 1966.

Блауберг И.В. Целостность и системность. – В кн.: Системные исследования. Ежегодник 1977. М.: Наука, 1977, с. 5-28.

Богданов В.В. Статистические концепции языка и речи. – В кн.: Статистика речи и автоматический анализ текста 1972. – Л.: Наука, 1973, с. 9-19.

Бодуэн де Куртенэ И.А. Избранные труды по общему языкознанию. Т. I-2. М., 1963.

Борода М.Г., Поликарпов А.А. Закон Ципфа-Мандельброта и единицы различных уровней организации текста. – Учен. зап. ТГУ, вып. 689. Тарту, 1984, с. 35-60.

Брушлинский А.В. Мышление и прогнозирование (логико-психологический анализ). – М.: Мысль, 1979. – 230 с.

БСЭ – Большая советская энциклопедия. Изд. 3-е. Т. I7. М., 1974.

Будагов Р.А. Система и антисистема в науке о языке. – ВЯ, 1978, № 4, с. 3-17.

Бычков В.Н. К проблеме обобщения и интерпретации ранговых распределений в статистической лингвистике. – Учен. зап. ТГУ, вып. 689. Тарту, 1984, с. 61-70.

Ванников Ю.В. Синтаксис речи и синтаксические особенности русской речи. – М.: Русский язык, 1979. – 296 с.

Венецкий И.Г., Кильдишев Г.С. Теория вероятностей и математическая статистика. Изд. 3-е. - М.: Статистика, 1975. - 264 с.
Вентцель Е.С. Исследование операций. - М.: Знание, 1976. - 64 с.
Вертель В.А., Вертель Е.В. Алгоритм получения частотного словаря с учетом длины словоформы. - В кн.: Статистика текста. Т. 2. Автоматическая переработка текста. Минск, 1970, с. 290-311.
Винер Н. Я - математик. / Перев. с англ. - М.: Наука, 1964.
Виноградов В.В. Современный русский язык. Вып. I, М., 1938.
Виноградов В.В. Русский язык (Грамматическое учение о слове). - М.; Л.: Учпедгиз, 1947. - 784 с.
Виноградов В.В. Стилистика. Теория поэтической речи. Поэтика. - М.: Изд-во АН СССР, 1967.
Винникова С.М. Выделение существительных и прилагательных при автоматическом анализе текста. - НТИ, сер. 2, 1976, № 3, с. 15-18.
Воронцак Е. Методы вычисления показателей лексического богатства текстов. - Семиотика и искусство. М., 1972, с. 232-249.
Гачечиладзе Т.Г., Цилосани Т.П. Об одном методе изучения статистической структуры текста. - В кн.: Статистика речи и автоматический анализ текста. Л.: Наука, 1971, с. 113-133.
Герд А.С. Основы научно-технической лексикографии. - Л.: ЛГУ, 1986. - 72 с.
Головин Б.Н. О роли статистики в описании языковых и речевых стилей. - Частотные словари и автоматическая переработка текстов. Тезисы докладов. Минск, 1968, с. 36-41.
Головин Б.Н. Язык и статистика. - М.: Просвещение, 1971. - 192 с.
Горькова В.И. Ранговое распределение на множествах научно-технической информации. - НТИ, сер. 2, 1969, № 7, с. 5-11.
Григорьева А.С. Статистическая структура русского эпистолярного текста (лексика частных писем). АКД, Л., 1981.
Гринберг Дж. Квантитативный подход к морфологической типологии языков. - Новое в лингвистике. Вып. 3. М., 1963, с. 60-94.
Дарчук Н.П. Индивидуальное и общее в лексической системе авторского стиля (на материале современной украинской художественной прозы). АКД. Киев, 1975.
Денисов П.Н., Костомаров В.Г. Стилистическая дифференциация лексики и проблема разговорной речи. - В кн.: Русская разговорная речь. Саратов, 1970, с. 69-75.
Денисов П.Н., Морковкин В.В., Сафьян Ю.А. Комплексный частотный словарь русской научной и технической лексики. - М.: Русский язык, 1978. - 408 с.
Добров Г.М. Прогнозирование науки и техники. - М.: Наука, 1969. - 208 с.
Дручков Н.А. Законы науки, их роль в познании. - М.: Знание, 1980. - 64 с.
Дюран Б., Одеял П. Кластерный анализ. / Перев. с англ. - М.: Статистика, 1977. - 128 с.
Енисеева И.И., Рукавишников В.О. Группировка, корреляция, распознавание образов. - М.: Статистика, 1977. - 144 с.
Ефремова Т.Ф. Из наблюдений над структурой соврем. русского языка на уровне морфов. - В кн.: Семантические и фонологические проблемы прикладной лингвистики. - М.: МГУ, 1968, с. 45-55.
Иван Л.И. Опыт статистического описания научно-технического стиля румынского языка. АКД. Л., 1966.
Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. / Перев. с англ. - М.: Мир, 1976. - 165 с.
Заде Л.А. Размытые множества и их применение в распознавании образов и кластер-анализе. - В кн.: Классификация и кластер. / Перев. с англ. - М.: Мир, 1980, с. 208-247.

Заплаткина Н.И. Графемная структура односложных слов в славянских языках. АКД. Киев, 1975.
Заплаткина Н.И. Системный подход к изучению языковых явлений. – В кн.: Система и структура языка в свете марксистско-ленинской методологии. Киев: Наукова думка, 1982, с. 25–42.
Засорина Л.Н. Автоматизация и статистика в лексикографии. – Л.: ЛГУ, 1966. – 128 с.
Захарова А.В. Опыт статистического исследования устной речи ребенка. – В кн.: Исследования по языку и фольклору. Вып. 2. Новосибирск, 1967, с. 16–38.
Зубов А.В. О вероятностно-алгоритмическом подходе к порождению текста. – В кн.: Экспериментальная фонетика и прикладная лингвистика. Минск, 1980, с. 178–183.
Калинин А.В. Лексика русского языка. – М.: МГУ, 1978. – 232 с.
Калинин В.М. Некоторые статистические законы математической лингвистики. – Проблемы кибернетики. Вып. II. М., 1964.
Калинин В.М. Функционалы, связанные с распределением Пуассона и статистическая структура текста. – Труды Математического института им. Стеклова. Вып. 79. М.; Л., 1965.
Калинина Е.А. Изучение лексико-статистических закономерностей на основе вероятностной модели. – В кн.: Статистика речи. Л.: Наука, 1968, с. 64–107.
Каширина М.Е. О типах распределения лексических единиц в тексте. – В кн.: Статистика речи и автоматический анализ текста, 1974. Л.: Наука, 1974, с. 335–360.
Клименко Н.Ф. Сложные глаголы в новогреческом и украинском языках. – В кн.: Структурная и математическая лингвистика. Вып. 2. Киев, 1974, с. 54–62.
Клочкова Э.А. О распределении классов слов в некоторых функциональных стилях русского языка. – В кн.: Вопросы славянского языкознания. Саратов, 1968, с. 109–118.
Клявина С.П. Сопоставление функциональных стилей латинского языка (лингвостатистическое исследование). АКД. Киев, 1977.
Кожина М.Н. Стилистика русского языка. – М.: Просвещение, 1977. – 224 с.
Козачков Л.С. Информационные системы с иерархической ("ранговой") структурой. – НТИ, сер. 2, 1978, № 2, с. 15–24.
Колмогоров А.Н. Теория вероятностей. – В кн.: Математика, ее содержание и значение. Т. 2. М., 1956.
Кондаков Н.И. Логический словарь. – М.: Наука, 1971. – 656 с.
Королев Э.И., Корсакова И.И., Сафронова М.В. Частота употребления слов в тексте и их лексические характеристики. – НТИ, сер. 2, 1984, № 2, с. 8–14.
Косериу Э. Синхрония, диахрония и история. / Перев. с исп. – В кн.: Новое в лингвистике. Вып. 3. М., 1963, с. 143–343.
Кохонен Т. Ассоциативная память. / Перев. с англ.–М.: Мир, 1980. – 240 с.
Кравец А.С. Природа вероятности (философские аспекты). – М.: Мысль, 1976.
Крускал Дж. Взаимосвязь между многомерным шкалированием и кластер-анализом. – В кн.: Классификация и кластер. / Перев. с англ. – М.: Мир, 1980, с. 20–41.
Крылов Ю.К. Об общей парадигме лингвостатистических распределений. – Учен. зап. ТГУ, вып. 628. Тарту, 1982, с. 80–102.
Крылов Ю.К. К вопросу о динамике нарастания объема словаря случайной выборки и связного текста. – Учен. зап. ТГУ, вып. 711. Тарту, 1985, с. 55–66.
Крылов Ю.К. Стационарная модель порождения связного текста. – Учен. зап. ТГУ, вып. 774. Тарту, 1987, с. 81–102.
Крылов Ю.К., Якубовская М.Д. Статистический анализ полисемии как языковой универсалии и проблема семантического тождества слова. – НТИ, сер. 2, 1977, № 3, с. 1–6.

Кубрякова Е.С. Морфологическая структура слова в современных германских языках. – В кн.: Морфологическая структура слова в индоевропейских языках. М.: Наука, 1970, с. 104–181.

Кульгав М.П. Имя существительное как основное средство воплощения "номинального/субстантивного стиля". – В кн.: Некоторые вопросы немецкой филологии. Пятигорск, 1971, с. 3–20.

Ланд К.Ч. Сравнительная статика в социологии. – В кн.: Математика в социологии. / Перев. с англ. М.: Мир, 1977, с.371–401.

Лебедев А.Н. Закономерности повторения слов в речи. – Психологический журнал, 1983, № 5, с. 11–22.

Лебедев А.Н. Нейрофизиологические пределы памяти человека и богатства его лексики. – Учен. зап. ТГУ, вып. 745. Тарту, 1986, с. 95–108.

Леонтьев А.А. Язык, речь, речевая деятельность. – М.: Просвещение, 1969. – 215 с.

Леонтьев А.А. Речевая деятельность. Проблемы математического моделирования речевой деятельности. – В кн.: Основы теории речевой деятельности. М.: Наука, 1974, с. 21–28, с. 73–80.

Лесохин М.М., Лукьяненков К.Ф., Пиотровский Р.Г. Введение в математическую лингвистику. – Минск: Наука и техника, 1982. – 264 с.

Лийв К., Тулдава Ю. О классификации текстов с помощью кластер-анализа. – Учен. зап. ТГУ, вып. 777. Тарту, 1987, с. 55–68.

Лукьяненков К.Ф., Нешитой В.В. Оценка степени связности слов в научно-техническом тексте. – В кн.: Вопросы лингвистики и методики преподавания иностранных языков. Минск, 1975, с.52–62.

Лурия А.Р. Язык и сознание. – М.: МГУ, 1979. – 320 с.

Маковский М.М. Системность и асистемность в языке. Опыт исследования антиномий в лексике и семантике. – М.: Наука, 1980. – 210 с.

Малаховский Л.В. Принципы частотной стратификации словарного состава языка. – В кн.: Статистика речи и автоматический анализ текста 1980. Л.: Наука, 1980, с. 99–105.

Мамедян Н.С. Об оценке параметров лингвистических распределений определенного класса. – Структурная и прикладная лингвистика. Вып. 3. – Л.: ЛГУ, 1987, с. 94–97.

Мартине А. Основы общей лингвистики. – В кн.: Новое в лингвистике. Вып. III. М.: Изд-во иностр. литературы, 1963, с. 347–566.

Мартыненко Г.Я. Некоторые статистические наблюдения на материале болгарского языка. – В кн.: Статистико-комбинаторное моделирование языков. М.; Л.: Наука, 1965, с. 327–339.

Мартыненко Г.Я. Некоторые закономерности концентрации и рассеяния элементов в лингвистических и других сложных системах. – В кн.: Структурная и прикладная лингвистика. Вып. I. Л.: ЛГУ, 1978, с. 63–79.

Мартыненко Г.Я. Типология лингвостатистических распределений. – Учен. зап. ТГУ, вып. 628. Тарту, 1982, с. 103–120.

Марусенко М.А. Об измерении связи отраслевых терминосистем с применением ЭВМ. – Учен. зап. ТГУ, вып. 591. Тарту, 1981, с. 74–81.

Марчакова И.В. Исследование частотного словаря ключевых слов. – НТИ, сер. 2, 1974, № 11, с. 7–15.

Маслов Ю.С. Введение в языкознание. – М.: Высшая школа, 1975. – 327 с.

Мистрик И. Математико-статистические методы в стилистике. – ВЯ, 1967, № 3, с. 42–52.

Митропольский А.К. Техника статистических вычислений. Изд. 2-е. – М.: Наука, 1971. – 576 с.

Мордвинцкая М.П., Слипченко Л.Н. Симметрия в лингвистических системах. – В кн.: Система и структура языка в свете марксистско-ленинской методологии. Киев: Наукова думка, 1982, с.70–84.

Налимов В.В. Вероятностная модель языка. О соотношении естественных и искусственных языков. Изд. 2-е. – М.: Наука, 1979. – 304 с.
Налимов В.В. Функция распределения вероятностей как способ задания размытых множеств. Наброски метатеории (дискуссия по работам Л. Заде). – Автоматика, 1979а, № 6, с. 80–87.
Налимов В.В., Мульченко З.М. Наукометрия. Изучение развития науки как информационного процесса. – М.: Наука, 1969. –192с.
Негуляев Г.А., Покрас Ю.Л., Колесников Л.И. Автоматизированный отбор лексики для информационно-поисковых тезаурусов. – НТИ, сер. 2, 1973, № 2, с. 16–24.
Нелюбин Л.Л. Перевод и прикладная лингвистика. – М.: Высшая школа, 1983. – 207 с.
Немченко В.Н. Современный русский язык. Словообразование. – М.: Высшая школа, 1984. – 256 с.
Нешитой В.В. Длина текста и объем словаря. Показатели лексического богатства текста. – В кн.: Методы изучения лексики. Минск: БГУ, 1975, с. 110–118.
Нешитой В.В. Система непрерывных распределений в информатике и лингвистике. – НТИ, сер. 2, 1984, № 3, с. 1–6.
Нешитой В.В. О взаимосвязи ранговых распределений со спектровыми. – НТИ, сер. 2, 1986, № 10, с. 19–25.
Никонов В.А. Длина слова. – ВЯ, 1978, № 6, с. 104–111.
Общее языкознание. Внутренняя структура языка. / Отв. ред. Б.А. Серебренников. – М.: Наука, 1972. – 565 с.
Общее языкознание. Методы лингвистических исследований. / Отв. ред. Б.А. Серебренников. – М.: Наука, 1973. – 318 с.
Овчинников Н.Ф. Принципы сохранения. М., 1966.
Ожегов С.И. Словарь русского языка. Изд. 5-е. М., 1963. – 900 с.
Орлов Ю.К. О статистической структуре сообщений, оптимальных для человеческого восприятия (к постановке вопроса). – НТИ, сер. 2, 1970, № 8, с. 11–16.
Орлов Ю.К. Обобщенный закон Ципфа-Мандельброта и частотные структуры информационных единиц различных уровней. – В кн.: Вычислительная лингвистика. М.: Наука, 1976, с. 179–202.
Орлов Ю.К. Модель частотной структуры лексики. – В кн.: Исследования в области вычислительной лингвистики и лингвостатистики. М.: МГУ, 1978, с. 59–118.
Орлов Ю.К. Статистическое моделирование речевых потоков. – Вопросы кибернетики. Вып. 41. М.; Л., 1978а, с. 66–99.
Орлов Ю.К. Информационные потоки: статистический анализ и прогнозирование. – НТИ, сер. 2, 1980, № 2, с. 23–30.
Панкрац Г.Я. Статистическое исследование фонологической структуры слова. – Учен. зап. ТГУ, вып. 591. Тарту, 1981, с. 82–90.
Панов Е.Н. Знаки, символы, языки. – М.: Знание, 1980. – 192 с.
Папп Ф. О некоторых количественных характеристиках словарного состава языка. – Slavica. T. 7. Debrecen, 1967, с.51–58.
Папп Ф. О машинной обработке одноязычных словарей (на материале венгерского языка). – НТИ, сер. 2, 1969, № 3, с. 20–29.
Папп Ф. Лингвостатистика и венгерский язык. – Учен. зап. ТГУ, вып. 518. Тарту, 1980, с. 15–37.
Перебейнос В.И. Определение надежности данных частотного словаря. – Учен. зап. ТГУ, вып. 689. Тарту, 1984, с. 103–110.
Перебийніс В.С. Кількісні та якісні характеристики системи фонем сучасної української літературної мови. Київ: Наукова думка, 1970. – 270 с.
Петренко Б.В. Исследование документального информационного потока на основе анализа запросов. – НТИ, сер. 2, 1974, № 10, с. 3–8.
Пиквер А. О применении дистрибутивно-статистического метода в морфемике. АКД. М., 1973.

Пиотровский Р.Г. Текст, машина, человек. – Л.: Наука, 1975. – 327 с.
Пиотровский Р.Г. Инженерная лингвистика и теория языка. – Л.: Наука, 1979. – 112 с.
Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А. Математическая лингвистика. – М.: Высшая школа, 1977. – 384 с.
Пиотровский Р.Г., Турыгина Л.А. Антиномия "язык – речь" и статистическая интерпретация нормы языка. – В кн.: Статистика речи и автоматический анализ текста. Л.: Наука, 1971, с.5–46.
Поликарпов А.А. Факторы и закономерности аналитизации языкового строя. АКД. М., 1976.
Поликарпов А.А. Полисемия: системно-квантитативные аспекты. – Учен. зап. ТГУ, вып. 774. Тарту, 1987, с. 135–154.
Прайс Д. Малая наука, большая наука. – В кн.: Наука о науке. М.: Прогресс, 1966, с. 281–384.
Пустыльник Е.И. Статистические методы анализа и обработки наблюдений. – М.: Наука, 1968. – 288 с.
Ракитов А.И. Философские проблемы науки. Системный подход.– М.: Мысль, 1977. – 270 с.
Рубашкин В.Ш. Признак и значение. – НТИ, сер. 2, 1976, № 3, с. 3–10.
Рузавин Г.И. Научная теория. Логико-методологический анализ. – М.: Мысль, 1978. – 246 с.
Рыбников К.А. Введение в методологию математики. – М.: МГУ, 1979. – 128 с.
Садовский В.Н. Основания общей теории систем.– М.: Наука, 1974. – 280 с.
Садовский В.Н. Развитие методологии системных исследований. – Общественные науки, 1979, № 3, с. 78–93.
Сачков Ю.В. Введение в вероятностный мир. Вопросы методологии. – М.: Наука, 1971. – 208 с.
Сводеш М. Лексикостатистическое датирование доисторических этнических контактов. – В кн.: Новое в лингвистике. Вып. I. – М.: Изд-во иностр. литературы, 1960, с. 23–52.
Симпченко Л.Д. Закономерности фонемной структуры слова в английском языке. – В кн.: Математическая лингвистика. Вып. I. Киев: КГУ, 1973, с. 104–109.
Солнцев В.М. Язык как системно-структурное образование. Изд. 2-е. – М.: Наука, 1977. – 341 с.
Соссюр Ф. де. Труды по языкознанию. /Перев. с франц. – М.: Прогресс, 1977. – 695 с.
Степанова М.Д., Хельбиг Г. Части речи и проблема валентности в современном немецком языке. – М.: Высшая школа, 1978. – 258 с.
Стивенс С.С. Экспериментальная психология. Т. I. / Перев. с англ. М., 1960.
Суппес П., Зинес Дж. Основы теории измерений. / Перев. с англ. – В кн.: Психологические измерения. М.: Мир, 1967, с. 9–100.
Суслов И.П. Вероятность в системе научных категорий. – В кн: Динамическая и вероятностная оптимизация экономики. Новосибирск, 1978, с. 43–58.
Тихонов А.Н. Система русского словообразования в свете количественных данных. – В кн.: Исследование деривационной подсистемы количественным методом. Владивосток, 1983, с. 61–73.
Тищенко В. Частота частин мови в різних функціональних стилях сучасної української мови.– В кн.: Питання структурної лексикології. Київ, 1970.
Тулдава Ю. Статистический метод сравнения лексического состава двух текстов. – Linguistica IV. Тарту, 1971, с.199–220.
Тулдава Ю.А. О статистической структуре текста. – В кн.: Советская педагогика и школа. Вып. 9. Тарту, 1974, с. 5–33.

Тулдава Ю.А. Опыт квантитативного анализа художественного стиля. – Учен. зап. ТГУ, вып. 396. Тарту, 1976, с. 122–141.
Тулдава Ю. О квантитативных характеристиках богатства лексического состава художественных текстов. – Учен. зап. ТГУ, вып. 437. Тарту, 1977, с. 159–175.
Тулдава Ю. Квантитативное исследование структуры односложного слова в эстонском языке. – Учен. зап. ТГУ, вып. 453. Тарту, 1978, с. 115–135.
Тулдава Ю. О некоторых квантитативно-системных характеристиках полисемии. – Учен. зап. ТГУ, вып. 502. Тарту, 1979, с. 107–141.
Тулдава Ю. К вопросу об аналитическом выражении связи между объемом словаря и объемом текста. – Учен. зап. ТГУ, вып. 549. Тарту, 1980, с. 113–144.
Тулдава Ю. О теоретико-методологических основах квантитативно-системного анализа лексики (2). – Учен. зап. ТГУ, вып. 585, Тарту, 1981, с. 114–133.
Тулдава Ю. Опыт классификации текстов с помощью кластер-анализа. – Учен. зап. ТГУ, вып. 591. Тарту, 1981а, с.136–157.
Тулдава Ю. Квантитативное исследование генетического состава лексики эстонского языка.-Учен. зап. ТГУ, вып. 628, Тарту, 1982, с. 136–166.
Тулдава Ю. Социальная дифференциация лексики эстонского языка с квантитативной точки зрения. – Учен. зап. ТГУ, вып. 658, Тарту, 1983, с. 149–177.
Тулдава Ю. Квантитативное исследование лексико-семантических групп в эстонском языке. – Учен. зап. ТГУ, вып. 656. Тарту, 1983а, с. 123–152.
Тулдава Ю.А. Проблемы и методы квантитативно-системного исследования лексики (на материале эстонского языка). ДД. Тарту, 1984.
Тулдава Ю.А. Развитие лексики эстонского языка по данным словарей XVII–XX вв. – Учен. зап. ТГУ, вып. 684. Тарту, 1984а, с. 115–126.
Тулдава Ю.А. Частотная структура текста и закон Ципфа. – Учен. зап. ТГУ, вып. 711. Тарту, 1985, с. 93–116.
Тулдава Ю.А. Длина слова и распределение слов по длине в тексте и словаре. – Учен. зап. ТГУ, вып. 736. Тарту, 1986, с. 150–166.
Тулдава Ю.А. О частотном спектре лексики текста. – Учен. зап. ТГУ, вып. 745. Тарту, 1986а, с. 139–162.
Тулдава Ю.А. О соотношении словоформ и лексем в тексте. – Владивосток (в печати).
Уфимцева А.А. Лексика. – Гл. IV в кн.: Общее языкознание. Внутренняя структура языка. М.: Наука, 1972, с. 394–455.
Филин Ф.П. О структуре современного русского литературного языка. – ВЯ, 1973, № 2, с. 5–12.
Филин Ф.П. Некоторые вопросы современного языкознания. – ВЯ, 1979, № 4, с. 19–28.
Философская энциклопедия. Т. I. – М.: Сов. энциклопедия, 1960. – 504 с.
Фрумкина Р.М. Применение статистических методов в языкознании. – ВЯ, 1960, № 4.
Фрумкина Р.М. К вопросу о так называемом законе Ципфа. – ВЯ, 1961, № 2.
Фрумкина Р.М. О вероятностном прогнозировании в речевом поведении. – Проблемы прикладной лингвистики. Тезисы межвузовской конференции. Ч. 2. М., 1969, с. 313–316.
Фукс В. Математическая теория словообразования. – В кн.: Теория передачи сообщений. М., 1957, с. 221–247.
Хайтун С.Д. Наукометрия – состояние и перспективы. – М.: Наука, 1983. – 344 с.

Хакулинен Л. Развитие и структура финского языка. Ч. I. Фонетика и морфология. – М.: Изд-во иностр. лит., 1953. – 312 с.
Херц М.М. О представительности текста заданной длины. – НТИ, сер. 2, 1969, № 6, с. 26–29.
Частотный словарь русского языка. / Под ред. Л.Н. Засориной. – М.: Русский язык, 1977. – 936 с.
Частотний словник сучасної української художньої прози. Том 1–2. – Київ: Наукова думка, 1981. – 864 + 854 с.
Чебанов С.Г. О подчинении укладов "индо-европейской" группы закону Пуассона. – Доклады АН СССР, новая серия, т.55, 1947, №2.
Четыркин Е.М. Статистические методы прогнозирования. – М.: Статистика, 1975. – 184 с.
Шайкевич А.Я. Опыт статистического выделения функциональных стилей. – ВЯ, 1968, № 1, с. 64–76.
Шайкевич А.Я. Дистрибутивно-статистический анализ текстов. АДД. Л., 1982.
Шептулин А.П. Категории диалектики как отражение закономерностей развития. – М.: Знание, 1980. – 64 с.
Шрейдер Ю.А. О возможности теоретического вывода статистических закономерностей текста (к обоснованию закона Ципфа). – В кн.: Проблемы передачи информации. Т.3, вып.1. М., 1967, с. 57–63.
Штофф В.А. Введение в методологию научного познания. – Л.: ЛГУ, 1972.
Эзремаа Р. Общая теория конструирования кластер-систем и алгоритмы для нахождения их численных представлений. – Труды Вычислительного центра. Вып. 42. Тарту, 1978, с. 53–77.
Яблонская Н.Н. Частотный словарь немецкого подъязыка хирургии. – Вопросы прикладной лингвистики. Вып. 6. Днепропетровск, 1976, с. 83–89.
Яблонский А.И. Структура и динамика современной науки (некоторые методологические проблемы). – В кн.: Системные исследования. Ежегодник 1976. М.: Наука, 1977, с. 66–90.
Якубайтис Т.А. Вероятностная характеристика слов с разным количеством слогов в латышском языке. – Изв. АН Латв. ССР, т. 7, 1963, с. 43–48.
Якубайтис Т.А. Части речи и типы текстов. – Рига: Зинатне, 1981. – 248 с.
Якубайтис Т.А., Скляревич А.Н. Вероятностные характеристики связных текстов. – Рига: АН ЛССР, 1978.
Ярцева В.Н. Количественные и качественные изменения в языке. – В кн.: Ленинизм и теоретические проблемы языкознания. М., 1970.
Alekseev P. Statistische Lexikographie. – Bochum: Brockmeyer, 1984.
Altmann G. Prolegomena to Menzerath's Law. – In: Glottometrika 2. – Bochum: Brockmeyer, 1980, pp. 1–10.
Altmann G. Statistik für Linguisten. / Quantitative Linguistik. Vol. 6. – Bochum: Brockmeyer, 1980a.
Altmann G., Buttler H.v., Rott W., Strauß U. A Law of Change in Language. – In: Historical Linguistics. – Bochum: Brockmeyer, 1983, pp. 104–115.
Antosch F. The Diagnosis of Literary Style with the Verb-Adjective Ratio. – In: Statistics and Style. New York, 1969, pp. 57–67.
Arapov M.V., Cherc M.M. Mathematische Modelle in der historischen Linguistik. – Bochum: Brockmeyer, 1983.
Bečka J.V. La structure lexicale des textes techniques en tchèque. – Philologica Pragensia, A. 15, 1972, No.1, p.25–32.
Bielfeldt H.H. Rucklaufiges Wörterbuch der russischen Sprache der Gegenwart. /2. Aufl. – Berlin: Akademie-Verlag, 1965.
Billmeier G. Über die Signifikanz von Auswahltexten. – In: Forschungsberichte des Instituts für deutsche Sprache, Nr. 2, Mannheim, 1968, S. 126–171.

Brookes B.C. Quantitative Analysis in the Humanities. - In: Studies on Zipf's Law. Bochum: Brockmeyer, 1982, pp. 65-115.
Busemann A. Stil und Charakter. Untersuchungen zur Psychologie der individuellen Redeform. Meisenheim (Glan), 1948.
Carroll J.B. On Sampling from a Lognormal Model of Word-Frequency Distribution. - In: Kučera H., Francis W.N. Computational Analysis of Present-Day American English. Providence, R. I, 1967, pp. 406-424.
Dugast D. La statistique lexicale. /Travaux de linguistique quantitative, 9. - Genève: Slatkine, 1980.
Embleton S.M. Statistics in Historical Linguistics. - Bochum: Brockmeyer, 1986.
Engwall G. Fréquence et distribution du vocabulaire dans un choix de romans français. - Stockholm: Skriptor, 1974.
Fickermann I., Markner B., Rothe U. Wortlänge und Bedeutungskomplexität. - In: Glottometrika 6. Bochum: Brockmeyer, 1984, S. 115-126.
Fucks W. Mathematical Theory of Word-Formation. - In: Information Theory. /Ed. by Colin Cherry. London, 1956, pp. 154-170.
Garcia Hoz V. Vocabulario usual, común y fundamental. Madrid, 1953.
Ginzburg R.S., Khidekel S.S., Knyazeva G.Y., Sankin A.A. A Course in Modern English Lexicology. - Moscow: Higher School Publishing House, 1966.
Guiraud, P. Les caractères statistiques du vocabulaire. Essai de méthodologie. Paris, 1954.
Guiraud P. Problèmes et méthodes de la statistique linguistique. Dordrecht, 1959.
Harlass G., Vater H. Zum aktuellen deutschen Wortschatz. - Tübingen: Gunter Narr Verlag, 1974.
Herdan G. Quantitative Linguistics. - London: Butterworths, 1964.
Herdan G. The Advanced Theory of Language as Choice and Chance. Berlin; Heidelberg; New York; Springer-Verlag, 1966.
Hoffmann L., Piotrowski R.G. Beitrage zur Sprachstatistik. - Leipzig: VEB Verlag Enzyklopädie, 1979.
Jiráková I. Зависимость количественного состава грамматических категорий порядка частей речи от объема частотных словарей русского языка. - Prague Studies in Mathematical Linguistics 5. Prague, 1976, pp. 37-52.
Juilland A., Brodin D., Davidovitch C. Frequency Dictionary of French Words. The Hague; Paris, 1970.
Köhler R. Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. - Bochum: Brockmeyer, 1986.
Krallmann D. Statistische Methoden in der stilistischen Textanalyse. Bonn, 1966.
Krámský J. A Quantitative Analysis of Italian Mono-, Di- and Trisyllabic Words. - Travaux linguistiques de Prague, 1. Prague, 1966, pp. 129-143.
Kučera H., Francis W.N. Computational Analysis of Present-Day American English. Providence, R. I. : Brown University Press, 1967.
Kuraszkiewicz W. Statystyczne badanie słownictwa polskich tekstów XVI wieku. - Z polskich studiów sławistycznych. Warszawa, 1958, str. 240-257.
Latviešu valodas biežuma vārdnīca. - Rīga; Zinātne, 1972.
Lounsbury F. Transitional Probability, Linguistic Structure and Systems of Habit-Family Hierarchies. - In: Psycholinguistics. A Survey of Theory and Research Problems. 2-nd ed. Bloomington, 1965, pp. 93-101.
Mandelbrot B. Structure formelle des textes et communication: deux études. - Word. Vol. 10, 1954, № 1, p. 1-27.

McKinnon A. Aberrant Frequency Words: Their Identification and Uses. - In: Glottometrika 2. Bochum: Brockmeyer, 1980, pp. 108-124.
Menzerath P. Die Architektonik des deutschen Wortschatzes. - Bonn: Dümmler, 1954.
Mistrík J. Frekvencia slov v slovenčine. Bratislava, 1969.
Muller Ch. Initiation à la statistique linguistique. - Paris: Larousse, 1968.
Muller Ch. Some Recent Contributions to Statistical Linguistics. - In: Statistical Methods in Linguistics 1976. Stockholm: Skriptor, 1976, pp. 136-147.
Müller W. Wortschatzumfang und Textlange. - Muttersprache, 81. Jg., Nº 4. Mannheim; Zürich, 1971, S. 266-276.
Pao M.L. Automatic Text Analysis Based on Transition Phenomena of Word Occurrences. - Journal of American Social Information Science. Vol. 29, 1978, pp. 121-124.
Piotrowski R.G. Text, Computer, Mensch. - Bochum: Brockmeyer, 1984.
Piotrowski R.G., Bektaev K.B., Piotrowskaja A.A. Mathematische Linguistik. - Bochum: Brockmeyer, 1985.
Ratkowsky D.A., Halstead M.H., Hantrais L. Measuring Vocabulary Richness in Literary Works: A New Proposal and a Re-Assessment of Some Earlier Measures. - In: Glottometrika 2. - Bochum: Brockmeyer, 1980, pp. 125-145.
Roberts A.H. A Statistical Linguistic Analysis of American English. - The Hague: Mouton, 1965.
Sambor J. Analiza stosunku "type-token". - Prace filologiczne. Tom XX. Warszawa, 1970, str. 65-70.
Sambor J. Menzerath's Law and the Polysemy of Words. - In: Glottometrika 6. Bochum: Brockmeyer, 1984, pp. 94-114.
Saukkonen P., Haipus M., Niemikorpi A., Sulkala H. Suomen kielen taajuussanasto. A Frequency Dictionary of Finnish. Porvoo; Helsinki: Juva, 1979.
Shannon C.E. A Mathematical Theory of Communication.- Bell System Technical Journal, vol. 27. 1948, pp. 379-423,623-656.
Simon H.A. On a Class of Skew Distribution Functions. - Biometrika. Vol. 42, 1955, pp. 425-440.
Somers H.H. Analyse mathématique du langage: Lois générales et mesures statistiques. Louvain, 1959.
Somers H.H. Statistical Methods in Literary Analysis. - In: The Computer and Literary Style. Kent, Ohio, 1966, pp.128-140.
Těšitelová M. On the So-Called Vocabulary Richness. - Prague Studies in Mathematical Linguistics, 3. Prague, 1972, pp. 103-120.
Tuldava J. Quantitative Relations between the Size of Text and the Size of Vocabulary. - SMIL Quarterly, Journal of Linguistic Calculus. Stockholm: Skriptor, 1977, Nº 4, pp.28-35.
Tuldava J. Sagedussõnastik leksikostatistilise uurimise objektina. - TRÜ Toimetised, vihik 413. Tartu, 1977, lk.141-171.
Viks Ü. Klassifikatoorne morfoloogia. Verb. Tallinn, 1980.
Villup A. A.H. Tammsaare romaani "Tõde ja õigus" I köite autori- ja tegelaskõne sagedussõnastik. - TRÜ Toimetised, vihik 446. Tartu, 1978, lk. 5-106.
Weibull W. A Statistical Theory of the Strength of Materials. Stockholm, 1939.
Zipf G.K. The Psycho-Biology of Language. Cambridge, Mass., 1935.
Zipf G.K. Human Behavior and the Principle of Least Effort. Cambridge, Mass.: Addison-Wesley Press, 1949.
Zsilka T. Stilisztika és statisztika. Budapest, 1974.
Õigekeelsussõnaraamat. / Toim. R. Kull, E. Raiet. - Tallinn: Valgus, 1976.

SUMMARY

The monograph "Problems and Methods of the Quantitative-Systemic Investigation of Vocabulary" by Juhan Tuldava (Tartu State University) presents a new synthetic approach to the study of vocabulary from the quantitative point of view.

In Chapter I some fundamental theoretical and methodological principles and concepts of the quantitative-systemic analysis of vocabulary are discussed, viz. the systemic character of vocabulary and the possibilities of representing the vocabulary and its subsystems as probabilistic systems characterized by the levels of stability (determination) and variability (chance). Classification and modelling of the material in the form of statistical distributions are regarded as the principal methods of description and interpretation of linguistic data in the framework of quantitative-systemic investigation. Due attention is paid to the combination of quantitative approach with qualitative analysis.

In Chapter II the main laws of the statistical organization of text and vocabulary are discussed. The frequency structure of text and vocabulary is considered to be a junction of two counterparts: the rank distribution and the spectral distribution of lexical units. The author gives a critical survey of some earlier and more recent versions of Zipf's law and discusses the problem of the theoretical substantiation of the law. A new approach to the analytical expression of the relation between the size of vocabulary and the size of text is proposed.

In Chapter III the phonetical, grammatical, and semantic aspects of the investigation of vocabulary from the quantitative-systemic point of view are considered. In Chapter IV special attention has been paid to quantitative analysis of the social differentiation of vocabulary. Some quantitative laws of diachronic lexicology are examined. Finally, the stylistic aspect of the functioning of vocabulary is discussed (the lexical "richness" of vocabulary, the measuring of the lexical proximity of texts, the cluster-analysis of texts on the basis of quantitative-lexical features).

The illustrative material has been taken from Estonian, Russian, English and other languages.

О Г Л А В Л Е Н И Е