Р. Г. ПИОТРОВСКИЙ, К. Б. БЕКТАЕВ,
А. А. ПИОТРОВСКАЯ

МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА

Р. Г. ПИОТРОВСКИЙ, К. Б. БЕКТАЕВ
А. А. ПИОТРОВСКАЯ

МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА

Рекомендовано к изданию
Министерством просвещения СССР
в качестве пособия
для студентов
педагогических институтов

МОСКВА «ВЫСШАЯ ШКОЛА» 1977

4
П 32
УДК 517+519.2:80(075)

Рецензенты: кафедра русского языка и общего языкознания Горьковского государственного университета и проф. А. М. Длин

Пиотровский Р. Г. и др.

П32 Математическая лингвистика. Учеб. пособие для пед. ин-тов. М., «Высш. школа», 1977.

383 с. с ил.

Перед загл. авт.: Р. Г. Пиотровский, К. Б. Бектаев, А. А. Пиотровская.

В книге рассматриваются различные вопросы языкознания, связанные с применением в нем математического анализа, теории вероятностей и математической статистики. Достаточное место уделено теоретическому обоснованию и приемам практического применения математических методов в изучении языка.

Предназначается для студентов педагогических институтов.

П $\frac{60602-222}{001(01)-77}$ 37—77

4

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

1. Язык и математика. В эпоху научно-технической революции математизация охватывает все сферы человеческой деятельности, в том числе и такие, казалось бы, чисто гуманитарные науки как языкознание. Проникновение математических методов в лингвистику обусловлено двумя причинами.

Во-первых, развитие языковедческой теории и практики требует введения все более точных и объективных методов для анализа языка и текста. Одновременно использование математических приемов при систематизации, измерении и обобщении лингвистического материала в сочетании с качественной интерпретацией результатов позволяет языковедам глубже проникнуть в тайны построения языка и образования текста.

Во-вторых, все расширяющиеся контакты языкознания с другими науками, например с акустикой, физиологией высшей нервной деятельности, кибернетикой и вычислительной техникой, могут осуществляться только при использовании математического языка, обладающего высокой степенью общности и универсальности для различных отраслей знаний. Особенно настойчиво математизируется языкознание в связи с использованием естественного языка в информационных и управленческих системах человек—машина—человек. В действующих системах машинного перевода, автоматического аннотирования, человеко-машинного диалога всякое сообщение на естественном языке перекодируется в математическом языке компьютера.

Говоря об особенностях взаимодействия языкознания и математики, следует иметь в виду, что как естественный язык, так и язык математики являются знаковыми (семиотическими) системами передачи информации.

Основные расхождения между этими языками связаны с различным построением языкового знака и знака математического.

Во-первых, лингвистический знак (слово, словосочетание, предложение) обычно включает в себя четыре компонента — *имя* (материальный носитель информации), *денотат* (отражение предмета из внешнего мира), *десигнат* (понятие о предмете) и *коннотат* (комплекс чувственно-оценочных оттенков, связанных с предметом и понятием о нем); знак математического языка включает только имя и десигнат (математическое понятие); сказанное иллюстрирует рис. 1.

Во-вторых, лингвистический знак многозначен; математический знак имеет, как правило, одно концептуальное значение.

В-третьих, лингвистический знак потенциально метафоричен, у знака математического метафоричность полностью отсутствует.

Все эти свойства лингвистического и математического знаков можно проследить, сравнив значения математического знака 7 и

слова *семерка*. Если 7 имеет единственное десигнативное математическое значение — «семь любых объектов», то слово *семерка* имеет несколько значений: «цифра 7», «карта в семь очков», «группа из семи человек» и т. п. При этом в значении слова *семерка* содержатся не только указанные десигнативные понятия, но оно может указывать на конкретный предмет, например на вполне определенную группу в семь человек. Одновременно это слово несет дополнительные коннотативные метафорические оттенки, связанные с такими словосочетаниями как «великолепная семерка», «семь чудес света», «семь смертных грехов», «семь дочерей Атланта (Плеяды)» и т. д.

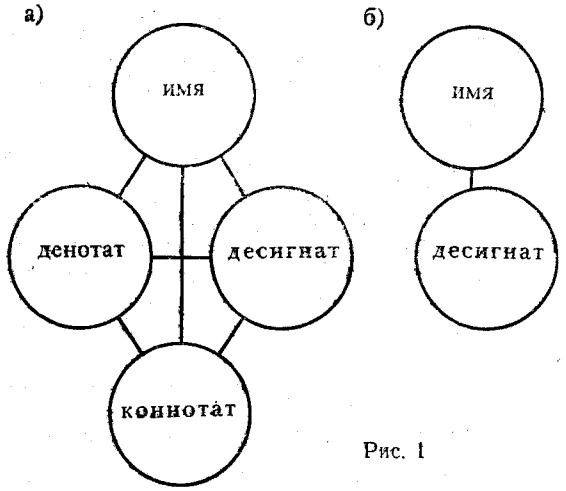


Рис. 1

Из всего сказанного вытекает еще одно важное различие между десигнативными значениями математического и лингвистического знаков.

Значение каждого математического знака легко представить в качестве множества элементов, причем такое множество имеет вполне четкие границы: значение знака 7 является множеством, охватывающим такие конкретные совокупности, которые включают только семь (не шесть и не восемь!) предметов.

Иначе организовано десигнативное значение лингвистического знака — оно также может рассматриваться как множество денотатов, однако это множество не всегда имеет четкие границы. Так, например, не удается определить смысловые границы слов *голубой* и *синий*, *голубой* и *зеленый*. Разные люди в зависимости от особенностей своего хроматического зрения будут называть показываемые им конкретные сине-голубые оттенки то синим, то голубым цветом. Нельзя также указать точную временну́ю границу, разделяющую значения слов *ночь* и *утро*. Иными словами, значения лингвистических знаков представляют собой нечеткие множества с размытыми границами [26, с. 207—214]; [65].

С многозначностью, метафоричностью и нечеткостью смысловых границ лингвистического знака связана также изменчивость его значения. В качестве примера снова возьмем русское прилагательное *голубой*. В 50-е годы это слово, судя по 3-му изданию «Словаря русского языка» С. И. Ожегова (М., 1957), имело в литературном русском языке только одно толкование: «с окраской светло-синего цвета». Однако, словарь-справочник, составленный по материалам прессы и литературы 60-х годов «Новые слова и значения» (М., 1971), указывает для слова *голубой* еще одно значение — «идеализированный», отмечая одновременно такие новые метафорические употребления как «голубое топливо», «голубой экран».

Особенности построения лингвистического языка приводят к тому, что естественный язык представляет собой нежестко организованную диффузную систему, которая воспринимается и используется человеком в значительной мере интуитивно.

Напротив, язык математики является хорошо организованной системой, существующей и функционирующей в виде логического построения, каждый элемент которого имеет осознанную значимость.

Конфронтация естественного языка и языка математики требует, чтобы каждому лингвистическому объекту был поставлен в соответствие некоторый математический объект. Лингвистический знак, например, словосочетание или слово и составляющие этот знак фигуры — фонемы, буквы, слоги — должны интерпретироваться с помощью знаков математических. Эта математическая интерпретация связана с расчленением лингвистического объекта и выделением в нем одного смыслового или сигнального компонента, который становится предметом дальнейшего исследования. Остальные сигнальные и смысловые элементы лингвистического объекта, а также разного рода метафорические коннотативные оттенки из рассмотрения исключаются.

Применение математических методов в языкознании имеет своей целью заменить обычно диффузную, интуитивно сформулированную и не имеющую полного решения лингвистическую задачу одной или несколькими более простыми, логически сформулированными и имеющими алгоритмическое решение математическими задачами. Такое расчленение сложной лингвистической проблемы на более простые алгоритмизуемые задачи мы будем называть *математической экспликацией* лингвистического объекта или явления.

Математическая экспликация интересна не только с чисто познавательной и теоретической точки зрения. Она совершенно необходима при решении прикладных вопросов, связанных с анализом и синтезом устной речи или информационной переработкой текста на ЭВМ. Математическая экспликация лингвистических объектов применяется не только при решении на ЭВМ несложных, хотя и трудоемких задач такого типа как составление частотных и алфавитных словников [3]; [8]; [22] или пословного и пообороотного машинного перевода [32 а, с. 286 и сл.], [32 б, с. 107—130], но также при составлении и реализации таких эвристических алгоритмов искус-

ственного интеллекта как семантический машинный перевод [32 в, с. 128—146] или тезаурусное реферирование текста [26, с., 248—268].

2. Комбинаторная и квантитативная лингвистика. Выбор математического аппарата в лингвистических исследованиях — вопрос не простой. Его решение зависит в первую очередь от того, как определяется предмет и основные понятия языкознания и его теоретического ядра — структурно-математической лингвистики.

Некоторые математики и лингвисты считают, что предметом математической и структурной лингвистики должно быть изучение грамматики, порождающей текст. При этом грамматика понимается как конечное множество детерминированных правил, в том числе неграмматических, а язык рассматривается как бесконечное число регулярных цепочек слов, порождаемых этой грамматикой. При этом подходе экспликация лингвистических объектов должна опираться на такие разделы «неколичественной» математики как теория множеств, математическая логика (в особенности, теории рекурсивных функций и бинарных отношений), теория алгоритмов и т. д.

Что же касается «количественных» разделов математики (математическая статистика, теория вероятностей, теория информации, математический анализ), то они считаются либо неприменимыми для экспликации лингвистических явлений, либо играющими вспомогательную роль. На основе применения «неколичественного», или как его иногда называют, «качественного» математического аппарата в теоретическом языкознании сформировалось направление, условно называемое комбинаторной лингвистикой. Это направление противопоставляется квантитативной (количественной) лингвистике [43, с. 273].

Методы детерминистского комбинаторного языкознания интенсивно разрабатываются в теории порождающих грамматик Хомского [45], в теоретико-множественных моделях Маркуса [56] и в других лингвистических направлениях.

Однако математическое языкознание не может ограничиться детерминистской, неколичественной экспликацией лингвистических объектов.

Во-первых, это ограничение затрудняет преобразование нечетких лингвистических множеств, элементы которых имеют вероятностные веса принадлежности, в четкие множества искусственных языков. Между тем указанное преобразование лежит в основе всех видов машинной переработки текста и автоматического распознавания смысла [26, с. 215—228].

Во-вторых, при таком ограничении вне сферы применения математических методов остается акустико-физиологическая и психолингвистическая проблематика речеобразования, а также стилистика и история языка, при изучении которых широко применяются не столько комбинаторные, сколько количественные измерения [18]; [21]; [27]; [32 в, с. 361—400].

Для того чтобы правильно оценить соотношение комбинаторных и количественных математических методов при описании языка и

текста, рассмотрим общую схему речевой деятельности и текстообразования.

Порождение текста определяется, с одной стороны, системой языка и ограничивающей ее действие нормой, а, с другой — совершенно независимой от языка внешней ситуацией (рис. 2).

Если согласиться с тем, что система языка есть механизм, порождающий тексты без каких-либо вероятностных ограничений [45]; [59], то станет ясным, что экспликация этой системы должна осуществляться с помощью тех неколичественных разделов математики, о которых мы говорили выше.

Рассматривая язык как неколичественную систему, комбинаторная лингвистика пытается описать механизм перехода от языка к речи с помощью тех же приемов «неколичественной» математики. Такие описания представляют собой контекстно-свободные грамматики, т. е. грамматики, не учитывающие контекстных ограничений на употребление отдельных лингвистических единиц и их сочетаний. В связи с этим контекстно-свободные грамматики порождают много цепочек, не являющихся реальными предложениями данного языка. Чтобы добиться порождения реальных текстов, необходимо перейти от контекстно-свободных грамматик к более сильным контекстно-зависимым грамматикам. Такие грамматики можно построить при условии, что к элементам системы языка применяются вероятностные оценки, а сам язык рассматривается как неколичественная производящая система, функционирование которой регулируется вероятностными ограничениями, заложенными в норме [32а, с. 5 — 46]; [47].

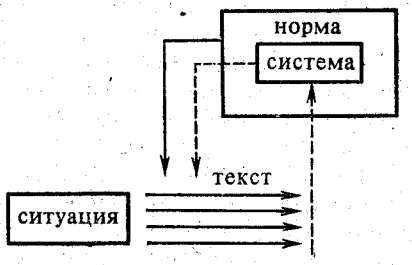


Рис. 2

Что же касается текста (речи), то он представляет собой линейную цепочку отграниченных друг от друга (дискретных) символов, (фонем, букв, слогов, слов). Каждый из символов встречается в тексте с определенной частотой и обладает особыми валентностями, т. е. лингвистическими способностями сочетаться с другими символами. Эти свойства лингвистических единиц в тексте эксплицируются в терминах теории вероятностей и математической статистики. К результатам вероятностно-статистического описания, взятым в сочетании с данными лингво-психологического эксперимента, может быть применен аппарат теории информации, с помощью которого удается количественно оценить как структурную организацию текста, так и заключенную в нем смысловую информацию.

Из всего сказанного следует, что математическая экспликация центральной проблемы современного языкознания «система языка — норма — текст» может быть осуществлена на основе применения методов как «качественной», так и «количественной» математики.

В связи с разработкой лингвистических аспектов искусствен-

ственного интеллекта как семантический машинный перевод [32 в, с. 128—146] или тезаурусное реферирование текста [26, с., 248—268].

2. Комбинаторная и квантитативная лингвистика. Выбор математического аппарата в лингвистических исследованиях — вопрос не простой. Его решение зависит в первую очередь от того, как определяется предмет и основные понятия языкознания и его теоретического ядра — структурно-математической лингвистики.

Некоторые математики и лингвисты считают, что предметом математической и структурной лингвистики должно быть изучение грамматики, порождающей текст. При этом грамматика понимается как конечное множество детерминированных правил, в том числе неграмматических, а язык рассматривается как бесконечное число регулярных цепочек слов, порождаемых этой грамматикой. При этом подходе экспликация лингвистических объектов должна опираться на такие разделы «неколичественной» математики как теория множеств, математическая логика (в особенности, теории рекурсивных функций и бинарных отношений), теория алгоритмов и т. д.

Что же касается «количественных» разделов математики (математическая статистика, теория вероятностей, теория информации, математический анализ), то они считаются либо неприменимыми для экспликации лингвистических явлений, либо играющими вспомогательную роль. На основе применения «неколичественного», или как его иногда называют, «качественного» математического аппарата в теоретическом языкознании сформировалось направление, условно называемое комбинаторной лингвистикой. Это направление противопоставляется квантитативной (количественной) лингвистике [43, с. 273].

Методы детерминистского комбинаторного языкознания интенсивно разрабатываются в теории порождающих грамматик Хомского [45], в теоретико-множественных моделях Маркуса [56] и в других лингвистических направлениях.

Однако математическое языкознание не может ограничиться детерминистской, неколичественной экспликацией лингвистических объектов.

Во-первых, это ограничение затрудняет преобразование нечетких лингвистических множеств, элементы которых имеют вероятностные веса принадлежности, в четкие множества искусственных языков. Между тем указанное преобразование лежит в основе всех видов машинной переработки текста и автоматического распознавания смысла [26, с. 215—228].

Во-вторых, при таком ограничении вне сферы применения математических методов остается акустико-физиологическая и психолингвистическая проблематика речеобразования, а также стилистика и история языка, при изучении которых широко применяются не столько комбинаторные, сколько количественные измерения [18]; [21]; [27]; [32 в, с. 361—400].

Для того чтобы правильно оценить соотношение комбинаторных и количественных математических методов при описании языка и

текста, рассмотрим общую схему речевой деятельности и текстообразования.

Порождение текста определяется, с одной стороны, системой языка и ограничивающей ее действие нормой, а, с другой — совершенно независимой от языка внешней ситуацией (рис. 2).

Если согласиться с тем, что система языка есть механизм, порождающий тексты без каких-либо вероятностных ограничений [45]; [59], то станет ясным, что экспликация этой системы должна осуществляться с помощью тех неколичественных разделов математики, о которых мы говорили выше.

Рассматривая язык как неколичественную систему, комбинаторная лингвистика пытается описать механизм перехода от языка к речи с помощью тех же приемов «неколичественной» математики. Такие описания представляют собой контекстно-свободные грамматики, т. е. грамматики, не учитывающие контекстных ограничений на употребление отдельных лингвистических единиц и их сочетаний. В связи с этим контекстно-свободные грамматики порождают много цепочек, не являющихся реальными предложениями данного языка. Чтобы добиться порождения реальных текстов, необходимо перейти от контекстно-свободных грамматик к более сильным контекстно-зависимым грамматикам. Такие грамматики можно построить при условии, что к элементам системы языка применяются вероятностные оценки, а сам язык рассматривается как неколичественная производящая система, функционирование которой регулируется вероятностными ограничениями, заложенными в норме [32а, с. 5 — 46]; [47].

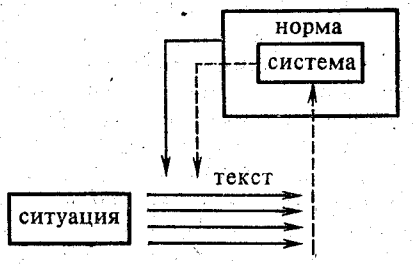


Рис. 2

Что же касается текста (речи), то он представляет собой линейную цепочку отграниченных друг от друга (дискретных) символов, (фонем, букв, слогов, слов). Каждый из символов встречается в тексте с определенной частотой и обладает особыми валентностями, т. е. лингвистическими способностями сочетаться с другими символами. Эти свойства лингвистических единиц в тексте эксплицируются в терминах теории вероятностей и математической статистики. К результатам вероятностно-статистического описания, взятым в сочетании с данными лингво-психологического эксперимента, может быть применен аппарат теории информации, с помощью которого удается количественно оценить как структурную организацию текста, так и заключенную в нем смысловую информацию.

Из всего сказанного следует, что математическая экспликация центральной проблемы современного языкознания «система языка — норма — текст» может быть осуществлена на основе применения методов как «качественной», так и «количественной» математики.

В связи с разработкой лингвистических аспектов искусствен-

ного интеллекта возникает необходимость формального описания внешних ситуаций, стимулирующих порождение текста. Для описания этих ситуаций используются как количественная, так и комбинаторная методика.

Что же касается моделирования непрерывных изменений языка во времени (диахроническая лингвистика), географическом пространстве (диалектология), а также в специально-профессиональном и художественном континууме (социолингвистика и стилистика), то целесообразно использовать понятия бесконечного множества, предельного перехода, непрерывности, т. е. понятия, составляющие основу математического анализа.

В области комбинаторной лингвистики наряду с фундаментальными исследованиями появилось уже немало работ типа учебных пособий, в которых систематизируются и популяризируются основные ее идеи [13]; [45]; [56]. В ином положении находится квантитативная лингвистика. Здесь можно указать лишь несколько книг и сборников, в которых исследуются или описываются отдельные вопросы приложения математического анализа, теории вероятностей и статистики в языкознании — см. [4]; [6]; [7]; [15].

Однако систематического изложения основных идей квантитативной лингвистики до сих пор нет. Предлагаемая читателю книга имеет своей целью восполнить этот пробел.

В первой части книги рассматриваются элементы математического анализа и их лингвистические приложения. С помощью этого аппарата строятся математические модели, описывающие: изменение лингвистических объектов во времени (гл. 1—4); распределение информации в письменном тексте (гл. 1, 2, 4), акустическую структуру устной речи (гл. 1).

Во второй части к лингвистическому материалу прилагается аппарат комбинаторики, теории вероятностей и математической статистики. Эта методика используется для: измерения смысловой информации слов и избыточности текста (гл. 5); описания функций распределения в тексте слогов, слов, словосочетаний и грамматических классов (гл. 6); построения статистических моделей текста и вероятностных характеристик норм языка (гл. 8, 9).

Математический аппарат, необходимый для построения всех этих моделей, чаще всего дается в виде определений без строгих математических доказательств, которые читатель всегда может найти в вузовских учебниках и пособиях по математическому анализу [28], теории вероятностей [10]; [14]; математической статистике [30]; [36] и лингво-статистике [6], [7].

Авторы приносят благодарность рецензентам проф. Б. Н. Головину и проф. А. С. Длину, а также доц. П. М. Алексееву и канд. техн. наук К. А. Разживину, замечания которых способствовали улучшению книги. Авторы благодарны В. В. Колесниковой, С. А. Моисеевой, П. В. Садчиковой и коллегам по группе «Статистика речи» за помощь при подготовке рукописи к печати. Кроме того, авторы выражают признательность редактору А. М. Суходскому, проделавшему большую работу по редактированию книги.

ЧАСТЬ ПЕРВАЯ

ИССЛЕДОВАНИЕ ЛИНГВИСТИЧЕСКИХ ПРОЦЕССОВ МЕТОДАМИ КВАНТИТАТИВНОЙ ЛИНГВИСТИКИ

ГЛАВА 1

ИСХОДНЫЕ ПОНЯТИЯ КВАНТИТАТИВНОЙ ЛИНГВИСТИКИ

§ 1. Множество лингвистических объектов

1. Понятие множества. Одно из основных понятий современной математики — понятие *множества*. Оно является первичным, т. е. не поддается определению через другие, более простые понятия. С понятием множества мы встречаемся довольно часто: буквы русского алфавита образуют множество, то же можно сказать о словоупотреблениях*, содержащихся в данном предложении, на данной странице и т. д.

Приведенные примеры обладают одним существенным свойством: все эти множества состоят из определенного конечного числа объектов, которые мы будем называть *элементами множества*. При этом каждый из объектов данного вида либо принадлежит, либо не принадлежит рассматриваемому множеству. Так, например, буква *ф* вне всякого сомнения принадлежит множеству букв, образующих русский алфавит, в то время как буква *f* этому множеству не принадлежит. Множества, включающие только такие объекты, принадлежность или непринадлежность которых к тому или иному множеству не вызывает сомнения, называются *четкими множествами*. Поскольку каждый рассматриваемый объект либо принадлежит, либо не принадлежит к рассматриваемому четкому множеству, эти множества всегда имеют ясно очерченные границы.

---

* В дальнейшем мы будем различать следующие лексикологические понятия: словоупотребление, форма слова (словоформа), слово, а также исходная форма слова. Под *словоупотреблением* понимается цепочка букв, заключенная между двумя пробелами в тексте и имеющая одно значение (омонимические словоупотребления рассматриваются как различные). Полностью совпадающие словоупотребления представляют одну *словоформу*. Слово выступает как некоторый класс (сумма) семантически и грамматически связанных между собой словоформ. Словоупотребление является единицей текста, слово — единицей двуязычного, толкового, энциклопедического и т. д. словаря. В этих словарях слово представлено в так называемой *исходной форме*, в качестве которой в русском языке выступает обычно именительный падеж единственного числа — для именных форм и инфинитив — для глагольных форм. Что же касается словоформы, то она используется обычно в качестве единицы частотного словаря.

Четким множествам противопоставлены *нечеткие* или «лингвистические» *множества*, включающие такие объекты, которые могут быть отнесены к тому или иному множеству лишь с определенной степенью достоверности. Понятие нечеткого множества проиллюстрируем на примере семантических полей прилагательных *младенческий, детский, отроческий, юношеский, молодой, среднего возраста, старый* [26, с. 210].

Чтобы определить границы семантических полей указанных слов и словосочетаний, произведем следующий эксперимент.

Предложим большой группе испытуемых — носителей русского языка относить к той или иной возрастной группе мужчин различного возраста. При этом выясняется, что интервал от 10 до 14 лет

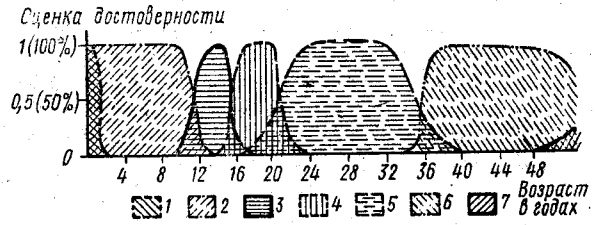


Рис. 3. Вероятностно-семантические поля русских прилагательных (и адъективальных словосочетаний) *младенческий (1), детский (2), отроческий (3), юношеский (4), молодой (5), среднего возраста (6), старый (старческий) (7)*

одними испытуемыми будет квалифицироваться как детский, а другими — как отроческий возраст. Аналогичным образом период от 17 до 23 лет может считаться либо как юношеский, либо как относящийся к молодому возрасту.

Если нанести результаты этого эксперимента на график, где на оси абсцисс отмечать конкретный возраст, а на оси ординат — процент достоверности его отнесения к той или иной понятийной области, то мы получим картину распределения семантических полей указанных терминов. При этом выясняется, что каждое из рассмотренных семантических полей представляет собой нечеткое подмножество с размытыми краями (рис. 3).

Объекты, попадающие на эти размытые края, относятся к указанным множествам лишь с известной долей достоверности. Так, например, двадцатилетний мужчина может быть с достоверностью 50% отнесен к множеству юношей, и с той же достоверностью — к множеству молодых людей.

Аппарат нечетких множеств начинает применяться для описания процессов мышления, лингвистических явлений и вообще для моделирования человеческого поведения, при котором допускаются частичные истины, а строгий математический формализм не является чем-то категорически необходимым [65, с. 6—12].

Множества, которые состоят из конечного числа элементов, называются *конечными множествами*.

К числу конечных множеств относится также и *пустое множество*, т. е. множество, не содержащее ни одного элемента. Введение понятия пустого множества связано с тем, что, определяя тем или иным способом множество, мы не можем знать заранее, содержит ли оно хотя бы один элемент. Например, множество двухбуквенных комбинаций *чы, бй, оъ*, можно считать пустым, если иметь в виду только русские тексты, написанные на литературном языке и не содержащие опечаток*.

Лингвистика чаще всего имеет дело с конечными множествами объектов. Однако приходится рассматривать и *бесконечные множества*. Например, бесконечным является множество всех словоупотреблений в текстах данного языка при условии, что этот язык беспрерывно порождает и будет порождать новые тексты без какого-либо ограничения во времени.

2. Способы задания множества. Существуют два различных способа задания множества. Можно дать полный перечень элементов этого множества. Этот способ называется *перечислением множества*. Элементы перечисляемого множества заключают обычно в фигурные скобки**. Например, множество A, состоящее из букв русского алфавита, вместе с пробелом (его обозначают знаком Δ) запишется так:

$$A = \{а, б, в, \ldots, ю, я, \Delta\}.$$

Другой способ состоит в том, что задается правило для определения того, принадлежит или не принадлежит любой данный объект рассматриваемому множеству. Этот способ называют *описанием множества*. При описании множеств используются различные символы, операции. Если A есть некоторое множество, а x — входящий в него объект, то символическая запись $x \in A$ означает, что x является элементом множества A; при этом говорят: «x входит в A», «x принадлежит A» (рис. 4, *а*).

Если x не принадлежит множеству A, то пишут $x \notin A$ (заштрихованная область на рис. 4, *б*). Пусть, например, A есть множество букв русского алфавита, а $л$ — буква этого алфавита; так как буква l в русский алфавит не входит, то можно записать $л \in A$, $l \notin A$.

* Однако этого нельзя утверждать с полной уверенностью относительно любого русского текста, поскольку эти комбинации могут появиться в записях диалектной речи, а также в результате орфографических ошибок или опечаток, например *чысто* вместо *чисто*, *оъ* вместо *об*. Кроме того, такие комбинации букв могут быть использованы в качестве каких-то условных обозначений, например, как обозначение серии перед номерами документов или денежных знаков. Поэтому множество двухбуквенных комбинаций *чы, бй, оъ* применительно к любому русскому тексту целесообразно рассматривать как нечеткое множество.

** Множества звуков, так же как и отдельные звуки, мы будем обозначать квадратными скобками [ ], а множества фонем и отдельные фонемы — косыми скобками / /. Для обозначения звуков и фонем используются знаки международной фонетической транскрипции [5, с. 475].

В том случае, когда речь идет о нечетком множестве, указывается степень достоверности, с которой x принадлежит множеству A. Это выражается записью $P(x \in A)$. Например, пусть A — множество юношей, а x обозначает двадцатилетнего мужчину; тогда, исходя из приведенных выше рассуждений, можно записать $0{,}5\ (x \in A)$.

Если имеются два множества A и B, причем каждый элемент множества A принадлежит множеству B, то множество A называется частью (или *подмножеством*) множества B. Записывается это так: $A \subset B$ или $B \supset A$. Соотношение, выраженное знаком $\subset$, называется *включением* (рис. 4, *в*).

Операцию включения можно проиллюстрировать на следующем лингвистическом примере. Русские [u] и [o], образующие множество огубленных (лабиализованных) гласных, принадлежат множеству гласных звуков. Таким образом, множество лабиализованных гласных следует рассматривать как подмножество, включенное во множество гласных звуков.

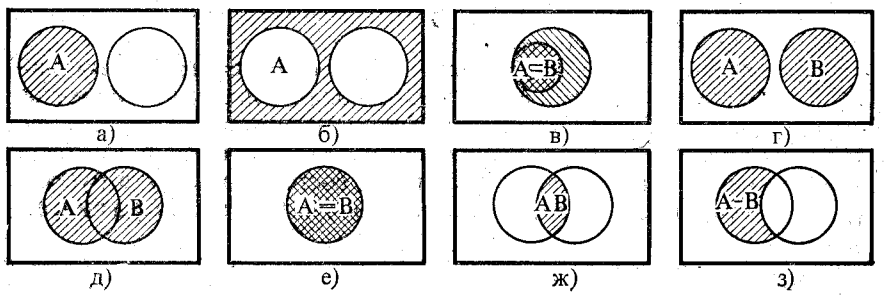


Рис. 4

3. Основные операции над множествами. Основными операциями, осуществляемыми над множествами, являются *сложение* (*объединение*), *умножение* (*пересечение*) и *вычитание*. Эти операции, как мы увидим дальше, не тождественны одноименным операциям, производимым над числами.

Объединением (или *суммой*) двух множеств называется множество, содержащее все такие и только такие элементы, которые являются элементами хотя бы одного из этих множеств (сумма множеств обозначается знаком $\cup$ или $+$). Это определение означает, что сложение множеств A и B есть объединение всех их элементов в одно множество $A + B$ или $A \cup B$ (рис. 4, *г*). Если одни и те же элементы содержатся в обоих множествах, то в сумму $A + B$ эти элементы входят только по одному разу (рис. 4, *д*). Так, если множество губных казахских согласных [p, b, m, w] есть A, а множество сонорных согласных [m, n, ń, w, l, r, j] есть B, то сумма $A + B$ состоит из элементов [p, b, m, n, ń, w, l, r, j]. Число элементов во множестве $A + B$ равно 9, а не 11, как это имело бы место при сложении чисел.

Сложение множеств, как и сложение чисел, обладает свойствами коммутативности: $A + B = B + A$, и ассоциативности: $(A + B) + C = A + (B + C)$, что легко проверить на примере множества казахских согласных.

Кроме того, сложение множеств обладает еще и такими свойствами, которые неприсущи сложению чисел; например, если $A \subset B$, то $A + B = B$. Действительно, если множество всех звонких согласных принять за A, а множество шумных согласных — за B, то сумма множеств A и B равна B, т. е. множеству шумных согласных. Всякое множество есть часть самого себя, т. е. $A \subset A$. Пустое множество есть часть всякого множества A.

Два множества A и B считаются *равными* ($A = B$), если они состоят из одних и тех же элементов, т. е. каждый элемент множества A является элементом множества B, и наоборот. Иначе говоря, $A \subset B$ и $B \subset A$ (рис. 4, *е*). Например, сравнивая множество A, состоящее из словоформ *вы, вас, вам, вами*, с множеством B, включающим формы склонения местоимения *вы*, убеждаемся, что $A \subset B$ и $B \supset A$, т. е. что оба множества равны.

Неравенство множеств A и B ($A \neq B$) указывает на то, что, по крайней мере, в одном из этих множеств есть такой элемент, которого нет в другом множестве. Например, множество ударных гласных звукотипов (фонем) по классификации Л. В. Щербы [40] не равно множеству тех же звукотипов (фонем) в классификационной схеме Р. И. Аванесова [1]; [2]. Легко заметить, что первое множество [a, e, i, o u, ы] содержит элемент [ы], которого нет во втором множестве [a, e, i, o, u]*.

Пересечением (или *умножением*) двух множеств A и B (обозначается $A \cap B$ или AB) называется множество тех элементов, которые принадлежат и к A, и к B (заштрихованная область на рис. 4, *ж*). Если мы обозначим множество ртовых чистых твердых смычных согласных [p, b, t, d, k, g] в русском языке через B, а множество заднеязычных твердых звуков [k, g, x] через A, то пересечение этих множеств AB или $A \cap B$ даст множество согласных [k, g].

Операция пересечения множеств обладает свойствами:

1) коммутативности: $AB = BA$;

2) ассоциативности: $(AB)C = A(BC)$;

3) дистрибутивности: $(A + B)C = AC + BC$.

Пересечение множеств обладает также такими свойствами, каких операция умножения чисел не имеет: например, если $A \subset B$, то $AB = A$ (см. рис. 4, *в*), в частности, $AA = A$. Эти свойства легко проверяются на множестве твердых смычных согласных.

* В действительности мы имеем здесь дело с разными разбиениями (группировками) одного и того же множества конкретных звуков — разбиениями, определяемыми разными фонологическими позициями ленинградской (Л. В. Щерба) и московской (Р. И. Аванесов) школ. Фонологическая позиция авторов настоящей работы изложена в [24].

Разностью двух множеств A и B называется множество всех таких элементов множества A, которые не содержатся во множестве B. Разность множеств обозначается $A - B$ или $A \setminus B$ (рис. 4, *з*).

Определение вычитания не требует, чтобы $A \subset B$. Если же $A \subset B$, то разность $B - A$ называется *дополнением* к множеству A во множестве B (см. рис. 4, *в*). Нетрудно видеть, что разность только что рассмотренных множеств согласных составляет

$$B \setminus A = [\text{p, b, t, d}].$$

Более сложное применение указанных операций для определения лингвистических понятий «язык», «диалект», «поддиалект», «говор», «подговор» см. в работе [25].

4. Упорядочение множества лингвистических объектов. В предыдущем параграфе мы рассматривали множества, не задаваясь вопросом о порядке расположения составляющих их лингвистических единиц. Однако порядок расположения единиц в том или ином лингвистическом множестве имеет принципиальное значение. Так, например, в толковых, энциклопедических словарях и разного вида справочниках слова расположены по алфавиту; размещение лексических единиц в другом порядке, например по убыванию частот, в корне меняет организацию этих множеств и их лингвистические приложения.

Рассматривая порядок размещения элементов внутри разного вида лингвистических множеств, мы приходим к понятию упорядоченного множества. Это понятие можно определить следующим образом: множество A называется *упорядоченным*, если для любых двух элементов один считается предшествующим другому.

Относительно любых элементов a_1, a_2, a_3 множества A это правило удовлетворяет следующим условиям: 1) если a_1 предшествует a_2, то a_2 не предшествует a_1 (а с и м м е т р и ч н о с т ь); 2) если a_1 предшествует a_2 и a_2 предшествует a_3, то и a_1 предшествует a_3 (т р а н з и т и в н о с т ь).

Одно и то же множество можно упорядочить многими различными способами — ср. в этом смысле разные построения словарей. Вместе с тем не для каждого множества удается задать конкретный и эффективный закон упорядочения. Например, неясно, как можно упорядочить множество значений всех слов или словосочетаний в произведениях Шекспира или Льва Толстого.

§ 2. Действительные числа

1. Понятие числа. Квантитативная лингвистика, исследующая количественную сторону языка и речи, постоянно оперирует не только понятием множества, но также и другим основным понятием математики — понятием *числа*.

Понятие числа выводится из понятий величины и измерения. Основное свойство величины состоит в том, что она может быть сопоставлена с другой определенной величиной того же класса, которая выступает в роли единицы меры. Сам процесс сопоставления первой величины с единицей меры и называется измерением. В итоге измерения мы получаем некоторое число, которое выражает отношение рассматриваемой величины к величине, принятой за единицу меры.

Если измеряемая величина соизмерима с единицей меры, то отношение между этой величиной и единицей меры выражается *рациональным* числом. К множеству рациональных чисел относятся числа целые и дробные (и те, и другие могут быть положительными и отрицательными), а также число нуль. Для лингвистов, разумеется, наиболее привычным является понятие *целого положительного (натурального) числа*: в каждом слове имеется целое положительное число букв, а в каждом предложении — целое положительное число слов и т. п. Упорядоченное множество целых положительных чисел 1, 2, 3, 4, ..., n составляет *натуральный ряд* чисел. Однако запаса одних лишь натуральных чисел оказывается недостаточным для квантитативных измерений текста. Так, например, для измерения средней встречаемости той или иной грамматической, лексической, фонологической единицы используются дробные числа. В некоторых зависимостях, описывающих лингвистические явления и процессы, используются отрицательные величины.

Если квантитативное языкознание ограничивало бы свои измерения четырьмя действиями элементарной математики, то запаса рациональных чисел было бы здесь вполне достаточно. Однако в лингвистике приходится решать задачи, которые требуют использования более сложных действий, например логарифмирования, извлечения корня. Бывает, что решение таких задач оказывается невозможным во множестве рациональных чисел. Так, например, располагая одними рациональными числами, мы не могли бы решить такое использующееся при исследовании информационного веса лингвистических единиц простейшее уравнение, с помощью которого оценивается нулевая энтропия английского алфавита:

$$H_0 = \log_2 27.$$

Действительно, среди рациональных чисел нельзя найти такое, которое будучи степенью числа 2, давало бы 27. Этим числом оказывается *иррациональное* число, которое изображается бесконечной десятичной дробью: $\log_2 27 = 4{,}7548\ldots$.

Запаса действительных (т. е. рациональных и иррациональных) чисел вполне достаточно для решения основных задач квантитативной лингвистики.

2. Множество действительных чисел. Все элементы множества *действительных чисел* — положительные и отрицательные (как рациональные, так и иррациональные, равно как целые и дробные), а также число нуль — упорядочены по величине. Это значит, что все эти числа связаны соотношениями взаимного расположения «равно» ($=$), «больше» ($>$), «меньше» ($<$). При этом для двух произвольных действительных чисел a и b имеет место одно и только

одно из трех соотношений* $a = b$, $a > b$, $a < b$. Короче говоря, два числа или два составленных из этих чисел выражения могут быть связаны отношениями равенства или неравенства.

Из курса средней школы известно, что каждому действительному числу соответствует определенная точка числовой оси, поэтому только что указанные алгебраические отношения можно перевести на язык геометрии. Так, например, записи $a = b$ эквивалентно предложение «точка a совпадает с точкой b», выражению $a > b$ соответствует высказывание «точка a лежит правее точки b», а вместо $a < b$ говорят: «a лежит левее b». Отсюда следует, что в тех случаях, когда между двумя лингвистическими элементами существуют отношения равенства и неравенства, они могут быть представлены не только алгебраически, но и геометрически.

§ 3. Лингвистическое явление как математическая величина

1. Математическая величина. При изучении количественных закономерностей языка приходится встречаться с такими лингвистическими явлениями, как употребительность слова или словосочетания и их порядок в частотном списке [3], длина звука, длина буквосочетания, информационный вес слога, морфемы или слова [23, с. 79—89], степень аналитичности языка [26, с. 190].

Если такое лингвистическое явление может быть выражено в виде числа, то его можно рассматривать в качестве математической величины.

2. Переменные и постоянные величины. Величина, которая при данном исследовании принимает различные значения, называется *переменной*, а величина, сохраняющая одно и то же значение,— *постоянной* (*константой*). Величины, которые в любых условиях сохраняют одно и то же числовое значение — так называемые *абсолютные постоянные* (например, отношение длины круга к диаметру, равное $\pi = 3{,}14159\ldots$) — встречаются довольно редко. Чаще мы будем иметь дело с величинами, сохраняющими одно и то же значение только при данных условиях исследования. Эти величины называются *параметрами*.

Понятия постоянной и переменной величин в значительной степени условны. Одна и та же величина может оказаться в одних условиях переменной, а в других — постоянной, и наоборот.

Рассмотрим, например, зависимость между частотой словоформы, которую она имеет в тексте длиной в N словоупотреблений, и ее номером в частотном словаре, составленном на основе данного текста. Эта зависимость выражается формулой (называемой обычно **законом Эсту—Ципфа—Мандельброта**), которая имеет следующий вид:

$$F_i = \frac{kN}{(i+\rho)^{\gamma}} = kN(i+\rho)^{-\gamma}. \qquad (1.1)$$

В этой зависимости F_i (частота словоформы) и i (номер ее в частотном словаре) выступают в качестве переменных величин, а величины N (длина исследованного текста), k (коэффициент относительной частоты наиболее частого слова), ρ (поправочный коэффициент частых слов) и γ (коэффициент лексического богатства текста) выступают в качестве параметров, сохраняющих постоянное числовое значение лишь для текста определенной длины, определенного стиля и тематики.

Зависимость Эсту—Ципфа—Мандельброта представляет собой весьма грубое приближение к истинной статистической структуре текста. Она более или менее удовлетворительно выполняется лишь для двух-трех тысяч наиболее частых словоформ.

Для описания статистически редких словоформ приходится оперировать другими зависимостями, в которых величины k и γ выступают уже в качестве переменных, а ρ может рассматриваться в качестве некоторого параметра [3]; [26, с. 106]; [55].

Переменная величина считается заданной, если указано множество значений, которое она может принимать. Это множество называется *областью изменения* переменной. Например, номер слова в списке может иметь только целочисленное значение, поэтому областью изменения переменной i в зависимости (1.1) является множество натуральных чисел.

Для обобщения некоторых формулировок и рассуждений бывает удобным рассматривать постоянную величину как частный случай переменной, у которой область изменения состоит из одного единственного числа. Так, например, в отдельных участках частотного списка область изменения величин k и γ можно с известным допущением охарактеризовать одним числом. Это дает нам право считать величины k и γ постоянными для определенных участков списка.

Геометрически можно изобразить область изменения переменной в виде некоторого множества точек числовой оси. Постоянной ве-

---

* Символы «>» «<» выражают так называемые *строгие неравенства*. Кроме того, теория неравенств оперирует отношениями *нестрогого неравенства* $a \geqslant b$ («a не меньше b», т. е. «a больше или равно b») и $a \leqslant b$ («a не больше b», т. е. «a меньше или равно b»). В дальнейшем нам придется иметь дело со следующими свойствами:

1) о б р а т и м о с т ь ю н е р а в е н с т в — если $a < b$, то $b > a$, если же $a > b$, то $b < a$;

2) о б р а т и м о с т ь ю р а в е н с т в — если $a = b$, то $b = a$;

3) т р а н з и т и в н о с т ь ю н е р а в е н с т в и р а в е н с т в — если $a < b$ и $b < c$, то $a < c$; аналогично, если $a = b$ и $b = c$, то $a = c$;

4) м о н о т о н н о с т ь ю с л о ж е н и я н е р а в е н с т в и р а в е н с т в — если $a < b$, а c — любое действительное число, то $a + c < b + c$; аналогично, если $a = b$, то $a + c = b + c$. Отсюда следует, что если к обеим частям неравенства прибавить любое действительное число, то получится новое неравенство того же смысла, т. е. *любой член неравенства можно перенести из одной части в другую с противоположным знаком*. Так, если имеется неравенство $a \leqslant b + c$, то, прибавив к обеим его частям $-c$, получим $a - c \leqslant b$;

5) м о н о т о н н о с т ь ю у м н о ж е н и я н е р а в е н с т в — если $a \geqslant b$ и $c \geqslant 0$, то $ac \geqslant bc$; если $a \geqslant b$ и $c < 0$, то $ac \leqslant bc$; иными словами, *при умножении (или делении) обеих частей неравенства на отрицательную величину знак неравенства меняется на противоположный*.

личине в этом случае соответствует множество, образованное одной точкой числовой оси.

3. Дискретные и непрерывные величины. Лингвистические явления могут быть представлены в виде *дискретных* и *непрерывных* переменных величин.

Область изменения дискретной величины состоит из отдельных изолированных точек числовой оси, например, —2, —1, 0, 1, 2. Область изменения непрерывной переменной величины состоит из всех точек, расположенных на каком-либо участке числовой оси, например между нулем и единицей, или из всех точек числовой оси, которые соответствуют всем действительным числам. Иными словами: между любыми двумя сколь угодно близкими значениями непрерывной величины может существовать любое количество ее промежуточных значений.

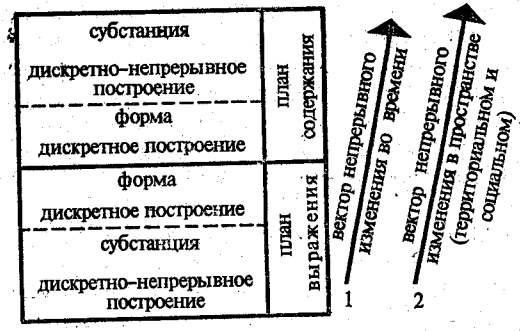


Рис. 5

4. Дискретность и непрерывность в языке и речи. Известно, что речь представляет собой последовательность дискретных, т. е. отграниченных друг от друга лингвистических единиц — фигур (букв, фонем, слогов) и знаков (морфем, слов, словосочетаний и даже предложений). Система языка выступает в виде сети отношений между единицами-инвариантами, также имеющими дискретную природу (фонемы, морфемы, слова, грамматические схемы). Отсюда иногда делается вывод, что в математической лингвистике, в том числе и в квантитативном языкознании, лингвистические объекты должны интерпретироваться с помощью дискретных величин и конечных множеств, а понятие непрерывности и связанное с ним понятие бесконечного множества для интерпретации лингвистических явлений малопригодны.

Чтобы оценить справедливость этого утверждения, обратимся к схеме речевой деятельности, изображенной на рис. 5.

Субстанция содержания охватывает здесь все то, что может быть предметом мысли. Таким образом, сюда входят как объекты, имеющие дискретную природу, так и понятия, которым присуща непрерывность (ср. цветовую гамму). *Структура (форма) содержания* представляет собой потенциально бесконечное множество идей. Эти идеи выступают в качестве особо упорядоченных в данном языке квазидискретных инвариантов плана содержания, образующих обычно нечеткие множества конкретных значений (см. § 1, п. 1).

Субстанция выражения представляет собой акустический, графический или иной материал, использующийся для формирования

знаков. Этот материал может иметь как дискретную природу (например, печатные буквы или импульсы тока, кодирующие эти буквы в ЭВМ), так и непрерывное строение (последовательности звуков, образующих устный текст). *Структура выражения* организует этот материал в систему дискретных и квазидискретных инвариантов (например, в систему фонем) плана выражения.

Если рассматривать процесс изменения языка во времени, а также территориальное и социальное его варьирование, то здесь преобладают непрерывные процессы. Особенно наглядно прослеживается это в области субстанции выражения. Действительно, утрата русских редуцированных гласных [ь] и [ъ] представляет собой с акустической точки зрения непрерывный процесс: их акустические следы на конце слова обнаруживаются даже в современном произношении [18, с. 33 и сл.]. Устранение произносительно-акустических различий между древнерусскими [ѣ] и [е] (обозначавшимся через «ять») тоже имело непрерывный характер не только в хронологическом, но и в территориальном отношении [1, с. 41 и сл.].

Что касается количественного измерения смысловой и статистической информации, содержащейся в тексте (субстанция плана содержания), то она также осуществляется с помощью непрерывных переменных величин (см. ниже, § 8).

Однако если аппарат непрерывной математики следует использовать для описания явлений и процессов, происходящих в субстанции выражения и субстанции содержания, которые рассматриваются многими лингвистами в качестве периферийных областей языка, то можно ли применять этот аппарат для интерпретации объектов структуры выражения и структуры содержания, составляющих структурное ядро языка и речи (ведь все объекты структурного ядра языка имеют дискретный характер)?

В ходе количественных группировок постоянно возникают ситуации, когда разность между смежными дискретными переменными, характеризующими эти группировки, очень мала по сравнению с их величинами. Так, например, словарь языка состоит из многих десятков и сотен тысяч лексических единиц. Увеличение словаря на одно вновь образованное или заимствованное слово несущественно с точки зрения общего объема словаря. Поэтому постоянно увеличивающийся со временем объем словаря можно рассматривать в качестве непрерывной переменной величины. В то же время меняющийся от варианта к варианту объем какого-либо древнего памятника не может рассматриваться в качестве непрерывной величины. Объем такого памятника обычно невелик, и его увеличение или уменьшение даже на одно слово представляет собой заметное изменение. Таким образом, изменение объема словаря памятника выступает в качестве дискретной переменной величины.

Непрерывность возникает также при наложении диалектных и особенно идиолектных (т. е. индивидуальных) систем в фонологии, грамматике и лексике [24, с. 32 и сл.]; [26].

Итак, для интерпретации лингвистических явлений могут быть использованы не только дискретные, но и непрерывные переменные величины. Последние должны использоваться в первую очередь при описании изменений языка во времени и пространстве, а также при экспликации информационной структуры текста.

§ 4. Понятие функции

1. Соответствия между лингвистическими множествами и функциональные зависимости. В предыдущих разделах нам уже несколько раз приходилось сравнивать разные лингвистические и нелингвистические множества. Сопоставление множеств приводит к понятию *соответствия*, которое подобно понятию множества является одним из основных понятий математической лингвистики.

Рассмотрим вопрос о соответствии множеств на примере соотношения индоевропейских и готских согласных.

Если опустить все спорные и дискуссионные вопросы, то индоевропейские звонкие смычные придыхательные согласные могут быть традиционно [57, с. 11—116] представлены в виде множества A, имеющего вид [bh, dh, gh], а соответствующие им готские согласные объединены во множество B, имеющее вид [b, d, g]. Между множествами A и B имеется соответствие, описываемое в германистике с помощью так называемого первого передвижения согласных [58, с. 37 и сл.]. Сущность этого соответствия состоит в том, что каждому звукотипу во множестве A взаимно однозначно соответствует определенный звукотип во множестве B. При этом переход от звукотипа множества A к звукотипу множества B совершается по четко определенному правилу: каждый из индоевропейских звукотипов теряет придыхательность.

Если два множества находятся в таком соответствии, что каждому объекту x множества A ставится в соответствие некоторый определенный объект y из множества B, то говорят, что между A и B существует *функциональная зависимость*. Правило f, по которому можно перейти от объекта $x \in A$ к соответствующему ему объекту $y \in B$, называется *функцией*. В нашем примере правило потери придыхательности у индоевропейских [bh, dh, gh] можно рассматривать как пример лингвистической функции.

Для обозначения функции используется запись $y = f(x)$.

2. Правила задания функций. Если соответствие между двумя множествами установлено, то принято говорить, что на множестве A задана функция со значениями, принадлежащими множеству B. Объекты x множества A, которым поставлены в соответствие элементы множества B, называются *значениями аргумента* (или *значениями независимой переменной*); в нашем случае в этой роли выступают индоевропейские звуки [bh, dh, gh]. Все множество A называется *множеством значений аргумента*, или *областью определения* (*существования*) *функции* $f(x)$.

Объекты y или $f(x)$ множества B называются *значениями функции* (или *значениями зависимой переменной*), соответствующими аргументу x; в нашем примере это готские [b, d, g]. Множество B объектов, поставленных в соответствие элементам множества A, называется *областью значений* (*изменения*) *функции* $f(x)$.

Вопрос о том, объекты какого множества следует считать независимыми переменными, решается обычно исходя из построения лингвистической задачи. Так, например, если первоначально заданными считать готские смычные звонкие согласные [b, d, g], от которых с помощью правила восстановления придыхательности можно реконструировать индоевропейские [bh, dh, gh], то первые можно рассматривать как значения аргумента, а вторые — как значения функции. Само же правило восстановления придыхательности выступает тогда в роли функции.

Введем еще одно важное понятие. Пусть задана функция $y = f(x)$, где x — независимая, а y — зависимая переменная. Если этого требуют интересы задачи, можно рассматривать y в качестве независимой переменной, а x — в качестве функции. В результате мы получаем новую функцию $x = \varphi(y)$, которая называется *обратной* по отношению к исходной (иначе — *прямой*) функции $y = f(x)$.

3. Алгоритм и вычислимые функции. С понятием функции тесно связано понятие *алгоритма*, которое обычно толкуется как точное предписание о выполнении в определенной последовательности некоторой системы операций для решения всех задач данного типа.

Если для рассматриваемой функции задан алгоритм, с помощью которого можно найти ее значения, то такая функция называется *вычислимой*.

Однако данное в п. 2 определение функции не предусматривает задания в правиле соответствия какого-либо алгоритма. Правило соответствия может быть каким угодно, в том числе и таким, которое не дает реальной возможности определять по значению аргумента соответствующее значение функции либо в силу того, что алгоритм для решения этой задачи в данный момент не найден, либо потому, что такой алгоритм не существует вовсе [38].

Таких «невычислимых» функций в лингвистике много. Рассмотрим, например, в качестве области определения функции упорядоченное по употребительности множество словоформ. Каждая словоформа из этого множества выступает в качестве значения независимой переменной, а несомая этой словоформой статистическая информация является значением функции. Правилом соответствия назовем утверждение о том, что более употребительные словоформы несут меньше информации, чем более редкие формы. Хотя функцию можно здесь считать заданной, алгоритм для определения значений этой функции отсутствует. Для того чтобы задать такой алгоритм, необходимо получить для каждой словоформы количественные оценки вероятности ее появления в тексте. Если же речь пойдет не о статистической, а о смысловой информации, то современное состояние наших знаний вообще не позволяет нам задать алгоритм для определения значений этой функции.

Для математической лингвистики и особенно для ее инженерных приложений вопрос вычислимости функции имеет принципиальное значение. Пока для функции не получен вычисляющий ее алгоритм, корректная переработка с ее помощью лингвистической информации, особенно на ЭВМ, практически невозможна.

§ 5. Числовые функции в лингвистике

1. Построение числовых функций. Изучение функциональных зависимостей предусматривает сопоставление множеств, различающихся не только по качественной природе, но и по количественной характеристике составляющих его объектов. Так, например, формула

$$p_i = k\,(i + \rho)^{-\gamma} \qquad (1.2)$$

описывает соответствие, существующее между множеством номеров i словоформ частотного списка и множеством их вероятностей p_i. Величина i выступает здесь в роли независимой переменной, а p_i является ее функцией. Алгоритм правила соответствия заключается в том, что к i прибавляется величина ρ, затем эта сумма возводится в степень $-\gamma$, а результат умножается на k. Аргумент i выражен числами натурального ряда и практически всегда ограничен, поскольку частотный словарь всегда содержит конечное число лексических единиц; параметры γ, ρ, k — действительные числа. Поэтому и область определения функции (1.2), и область ее значений являются конечными множествами. Если же представить себе частотный список с бесконечным числом лексических единиц, то аргумент i выражался бы числом, лежащим в промежутке $(1, +\infty)$. Тогда и область определения функции, и область значений были бы бесконечными множествами.

В только что рассмотренной функциональной зависимости значения аргумента и значения функции — числа, а область определения и область значений функции — числовые множества. Такие зависимости называются *числовыми функциями числового аргумента*, или сокращенно *числовыми функциями*. Наряду с числовыми функциями от числового аргумента существуют функции с нечисловыми или произвольными областями определения (а также с нечисловыми областями значений). Примером таких функций служит зависимость между индоевропейскими звонкими смычными придыхательными согласными и их готскими производными.

Квантитативная лингвистика имеет обычно дело с числовыми функциями.

2. Способы задания числовых лингвистических функций. Числовая функция, описывающая лингвистические процессы, может задаваться либо таблицей, либо в виде формулы, либо с помощью графика.

В современных эмпирических языковедческих исследованиях численное соответствие между значениями аргумента и функции устанавливается с помощью наблюдения. Поэтому здесь обычно ис-

пользуется табличный способ задания функции. Основное достоинство таблицы состоит в том, что она непосредственно, без всяких дополнительных измерений и вычислений соотносит значения аргумента и отвечающие им значения функции. Однако этот способ имеет существенный недостаток, состоящий в том, что таблица дает лишь выборочные значения среди всех значений, принимаемых аргументом и функцией. При этом в таблице может не оказаться тех значений аргумента и функций, которые для нас как раз и представляют наибольший интерес. Кроме того, табличный способ имеет слабую наглядность: по таблице трудно бывает определить характер изменения функции в зависимости от изменения аргумента.

Аналитический (формальный) способ дает возможность вычислять значения функции при любом значении аргумента. Одновременно этот способ позволяет анализировать с помощью математического аппарата различные свойства функции, в том числе и такие, которые невозможно изучать методом прямого наблюдения. Недостатком аналитического способа является отсутствие наглядности.

Наглядное представление функции дает графический способ. Его преимущество заключается в том, что он дает возможность охватить рассматриваемую функциональную зависимость «одним взглядом» и быстро выявить основной характер функции.

При решении практических задач квантитативной лингвистики к этому способу представления функции обращаются в том случае, когда в распоряжении исследователя имеются лишь ряды опытных данных. Один из этих рядов рассматривается в качестве значений аргумента, а другой — в качестве значений функции. Тогда по табличным данным строится график (кривая) зависимости, а по виду графика подбирается уравнение, соответствующее предполагаемой природе исследуемого лингвистического явления.

Составить формулу, описывающую лингвистическое явление, минуя табличный и графический способы представления функции, можно лишь тогда, когда заранее известен внутренний механизм этого явления. При моделировании диахронических процессов и описании построения устной и письменной речи подобная ситуация встречается сравнительно редко. Чаще приходится иметь дело с задачами такого типа, когда по экспериментальным данным, обобщенным в таблице, строится кривая, а по виду кривой определяется соответствующее ей уравнение.

Чтобы избежать ошибок при переходе от словесного лингвистического описания явления и эмпирических табличных данных к графику и уравнению, лингвист всегда должен иметь перед глазами набор таких числовых функций и соответствующих им графиков, с помощью которых обычно моделируются лингвистические процессы и явления. Поэтому каждому разделу, посвященному определенному виду лингвистического моделирования, предпослано краткое описание того математического аппарата, который обычно используется в этом моделировании, а также будет необходим в последующих разделах.

§ 6. Элементарные функции

Диахронические процессы, построение устной и письменной речи, а также диалектно-стилевую вариантивность можно моделировать в первом приближении с помощью различных *элементарных функций*. Рассмотрим некоторые из них.

1. Полиномы. *Полином* (*многочлен*) n-й степени задается в общем виде выражением

$$y = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n, \qquad (1.3)$$

где n — натуральное число, a_0, a_1, ..., a_n — постоянные действительные числа (коэффициенты).

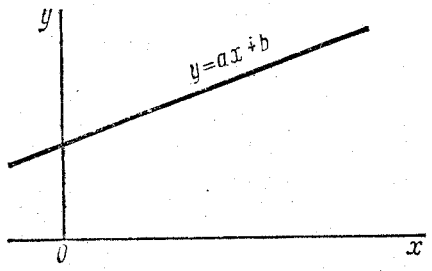


Рис. 6

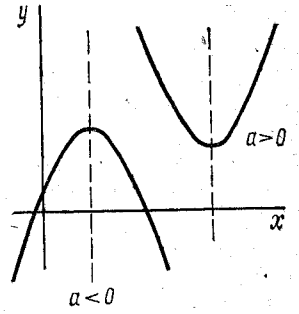


Рис. 7

Простейшими видами полинома являются следующие функции.

1. *Двучлен первой степени* (*линейная функция*)

$$y = ax + b. \qquad (1.4)$$

График этой функции — прямая линия. Постоянная величина a представляет угловой коэффициент прямой, равный тангенсу угла α, образованного ею с осью Ox, а коэффициент b — отрезок, отсекаемый ею на оси Oy (рис. 6). Областью определения функции является вся ось Ox (т. е. $-\infty < x < \infty$).

2. *Трехчлен второй степени* (*квадратичная функция*)

$$y = ax^2 + bx + c, \qquad (1.5)$$

графиком которой служит парабола с вертикальной осью симметрии $x = -b/(2a)$. При $a > 0$ функция сначала убывает и, достигнув минимума, снова возрастает; при $a < 0$ сначала возрастает, а достигнув максимума, начинает убывать (рис. 7).

3. *Степенная функция*

$$y = ax^n \qquad (1.6)$$

представляет собой частный случай полинома n-й степени, когда коэффициенты a_1, a_2, ..., a_n равны нулю. При целых и положительных значениях x функция определена на всей оси Ox. На рис. 8

изображены графики функций $y = x^n$ при различных значениях параметра n.

2. Дробно-рациональная функция. *Дробно-рациональная функция*, представляющая собой отношение двух многочленов, определяется равенством

$$y = \frac{a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n}{b_0x^n + b_1x^{n-1} + \ldots + b_{n-1}x + b_n}. \qquad (1.7)$$

Здесь n — натуральное число, a_0, a_1, ..., a_n и b_0, b_1, ..., b_n — коэффициенты.

Простейшим случаем дробно-рациональной функции является зависимость

$$y = a/x. \qquad (1.8)$$

Эта функция.* выражает закон обратной пропорциональности между переменными x и y.

Так как аргумент может принимать любые значения, кроме нуля (деление на нуль невозможно), то значение функции при $x = 0$ не существует, и кривая распадается на две не соединенные между собой ветви и, как говорят, терпит разрыв (ср. гл. 3, § 3, п. 1).

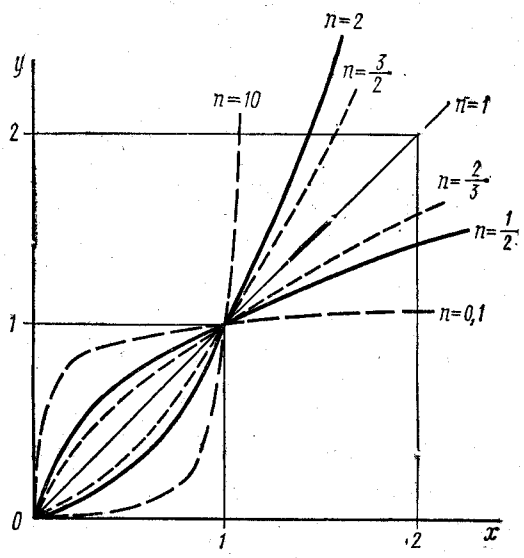


Рис. 8

Областью существования функции и областью значений функции являются два интервала $(-\infty, 0)$, $(0, +\infty)$. Графиком функции (1.8) служит равносторонняя гипербола, ветви которой асимптотически приближаются к осям координат. Если $a > 0$, то ветви гиперболы лежат в I и III четвертях; если же $a < 0$, то ветви расположены во II и IV четвертях (рис. 9).

3. Показательная функция. *Показательная* (экспоненциальная) *функция* задается равенством

$$y = a^x, \qquad (1.9)$$

в котором основание a есть постоянное положительное число, не равное единице. Графиком этой функции служит монотонно возрастающая или убывающая кривая (экспонента), которая асимптотически приближается к оси Ox (рис. 10 и 11). Область определения

* Ее можно рассматривать и как степенную функцию вида $y = ax^{-1}$.

функции — вся ось абсцисс, т. е. $-\infty < x < \infty$, область ее знаний $0 < y < \infty$.

В лингвистических приложениях используется показательная функция, где в качестве постоянной a взято *число Эйлера* $e = 2{,}718\ldots$, выступающее в качестве основания натуральных логарифмов. Функция

$$y = e^x \tag{1.10}$$

называется *экспоненциальной*, а ее график — экспонентой.

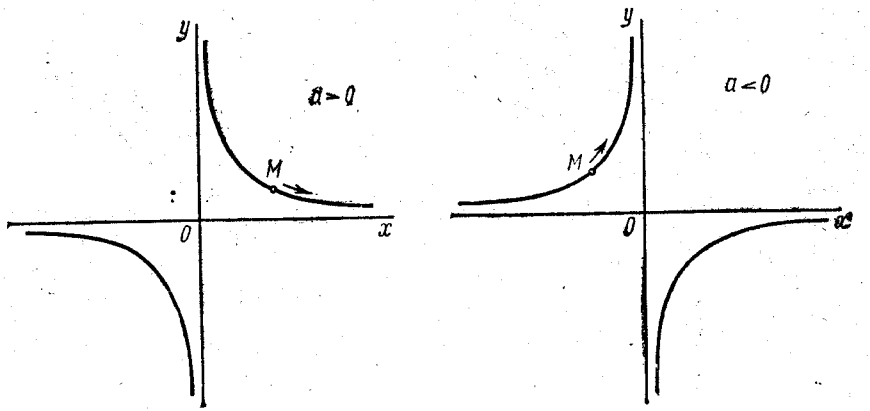


Рис. 9

4. Логарифмическая функция. *Логарифмической* функцией называется функция вида

$$y = \log_a x, \tag{1.11}$$

где основание логарифма a есть положительное, отличное от единицы число. По определению логарифма, равенство $y = \log_a x$ является обратным по отношению к равенству $x = a^y$, и, наоборот, $y = a^x$ обратно $x = \log_a y$. Поэтому график логарифмической функции (логарифмика) является зеркальным отображением экспоненты относительно биссектрисы I координатного угла (см. рис. 10 и 11). Логарифмика асимптотически приближается к оси Oy. Областью определения функции служит правая полуось Ox, т. е. $0 < x < \infty$, а областью значений функции — вся ось Oy, т. е. $-\infty < y < \infty$.

5. Периодические функции. Для ряда лингвистических явлений характерна отчетливо выраженная периодичность. Такой периодичностью обладают не только акустическая субстанция речевого сигнала, но также и распределение информации в письменном тексте. Явления этого типа описываются с помощью *периодических функций*.

Основная особенность периодических функций состоит в том, что имеется такое постоянное число T, прибавление которого к любому допустимому значению аргумента не изменяет значения функции, т. е.

$$y = f(x + T) = f(x). \tag{1.12}$$

Наименьшее положительное число τ, от прибавления которого к любому допустимому значению аргумента значение функции не изменяется, называют *периодом*. Следует иметь в виду, что $f(x + n\tau) = f(x)$, где n — любое число.

К периодическим функциям относятся тригонометрические и обратные тригонометрические функции.

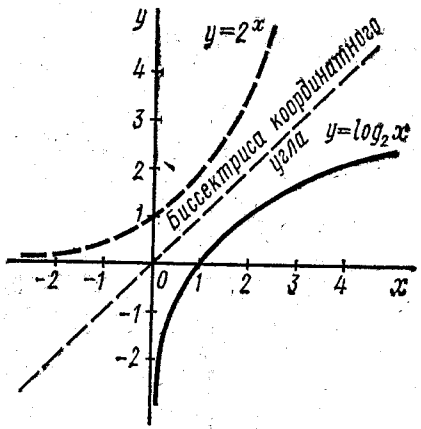


Рис. 10

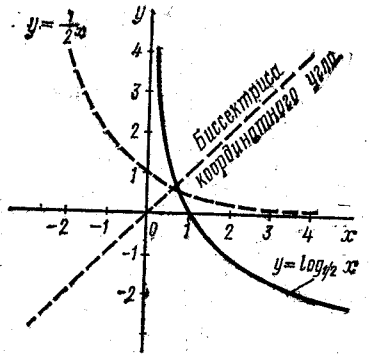


Рис. 11

6. Тригонометрические функции. Рассмотрим следующие четыре основные *тригонометрические функции*.

1. *Синус*:

$$y = r \sin(\omega x + \varphi); \tag{1.13}$$

здесь и далее x — аргумент, выраженный в радианной мере*, r, ω, φ — коэффициенты; если $r = \omega = 1$ и $\varphi = 0$, то получается функция

$$y = \sin x, \tag{1.14}$$

графиком которой является обыкновенная синусоида с периодом $T = 2\pi$ (рис. 12), причем $-\infty < x < \infty$, $-1 \leqslant y \leqslant 1$.

2. *Косинус*:

$$y = r\cos(\omega x + \varphi) = r\sin\left(\omega x + \varphi + \frac{\pi}{2}\right); \tag{1.15}$$

его графиком служит косинусоида, также имеющая период $T = 2\pi$ и сдвинутая относительно синусоиды (1.13) на $\pi/2$, причем $-\infty < x < \infty$, $-1 \leqslant y \leqslant 1$.

3. *Тангенс*:

$$y = \operatorname{tg} x; \tag{1.16}$$

* Единицей радианного измерения служит дуговой радиан, представляющий собой отношение дуги, равной по длине радиусу, к самому радиусу. Угол, образованный этой дугой, составляет $57°17'45''$. Отсюда $\sin 1 = \sin 57°17'45'' = 0{,}8414$.

график этой функции (тангенсоида) состоит из ряда отдельных одинаковых бесконечных ветвей, каждая из которых размещается в вертикальной полосе шириной (периодом) в $T = \pi$ (рис. 13); область существования основных значений функции лежит в интервале $(-\pi/2, \pi/2)$, а область значений функции — в интервале $(-\infty, \infty)$.

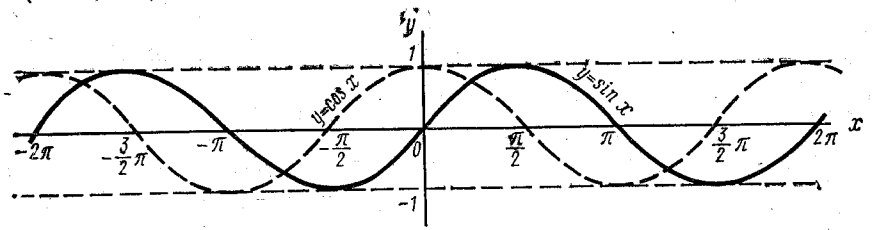


Рис. 12

4. *Котангенс*:

$$y = \operatorname{ctg} x; \tag{1.17}$$

графиком этой функции служит котангенсоида, являющаяся зеркальным отображением тангенсоиды относительно оси Oy (рис. 13). Область существования основных значений функции лежит в интервале $(0, \pi)$, а область значений функции — в интервале $(-\infty, \infty)$.

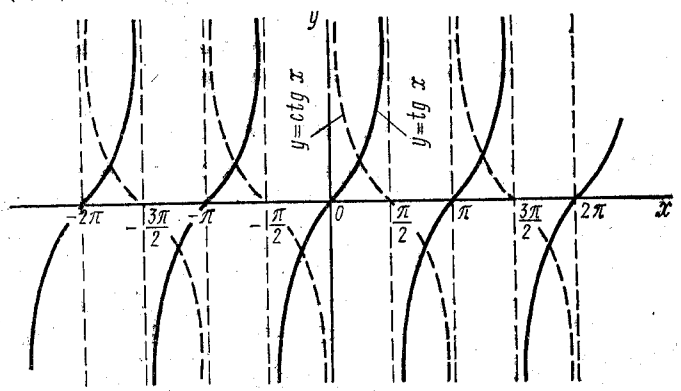


Рис. 13

7. Обратные тригонометрические функции. Если рассматривать в качестве аргумента значения синуса, косинуса, тангенса или котангенса, а функцией считать соответствующие этим аргументам значения угла или дуги, то мы получаем *обратные тригонометрические функции*.

1. *Арксинус*:

$$y = \arcsin x, \tag{1.18}$$

область определения этой функции есть интервал $[-1, 1]$, а область ее значений лежит в интервале $[-\pi/2, \pi/2]$.

2. *Арккосинус*:

$$y = \arccos x, \tag{1.19}$$

эта функция определена в интервале $[-1, 1]$, ее значения заключены в интервале $[0, \pi]$.

3. *Арктангенс*:

$$y = \operatorname{arctg} x, \tag{1.20}$$

областью существования функции (1.20) является вся числовая ось, т. е. $-\infty < x < \infty$, основные значения функции лежат в интервале $(-\pi/2, \pi/2)$.

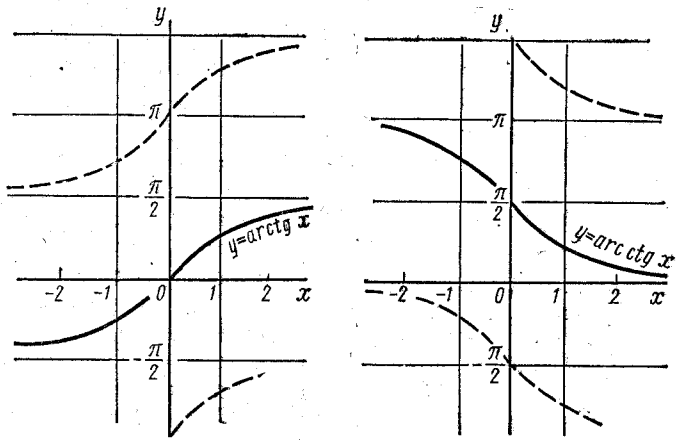


Рис. 14

4. *Арккотангенс*:

$$y = \operatorname{arcctg} x, \tag{1.21}$$

областью определения функции (1.21) также является вся числовая ось, т. е. $-\infty < x < \infty$, а основные значения функции лежат в интервале $(0, \pi)$.

Графики обратных тригонометрических функций получаются из графиков соответствующих тригонометрических функций с помощью зеркального отображения последних относительно биссектрисы I координатного угла. Это легко можно проверить, сравнив графики функций $y = \operatorname{tg} x$ и $y = \operatorname{ctg} x$ (рис. 13) с графиками соответствующих им обратных функций $y = \operatorname{arctg} x$ и $y = \operatorname{arcctg} x$ (рис. 14).

§ 7. Диахронический скачок и его моделирование с помощью элементарных функций

В современных работах по диахронической лингвистике часто приводятся таблицы, включающие количественные данные об употребительности исследуемого явления на различных этапах его истории. Цель таких таблиц состоит в том, чтобы показать динамику

лингвистического процесса. Но, как уже говорилось, таблица обладает слабой наглядностью и может не включать в себя тех значений аргумента и функции, которые представляют наибольший интерес с точки зрения динамики исследуемого процесса. Поэтому для более углубленного анализа лингвистического процесса следует, используя табличные данные, дать графическое представление, а затем и аналитическую модель исследуемого процесса.

1. Развитие нулевых форм родительного падежа множественного числа у существительных — единиц измерения в русском языке XIX—XX в. Рассмотрим в этой связи историю развития форм родительного падежа множественного числа у русских единиц измерения типа *вольт*, *рентген*, *радиан*. В русских научно-технических текстах XIX в. употреблялись регулярные образования родительного падежа множественного числа: *вольтов*, *рентгенов*. Однако, начиная с конца 80-х годов, отмечается нарастающее употребление необычных форм: *вольт*(ъ), *рентген*(ъ), *радиан*(ъ), совпадающих с именительным падежом единственного числа. Через 20—30 лет необычные формы утверждаются не только в профессиональной речи, но и в литературном языке. Так сформировалась новая лексико-грамматическая группа имен существительных, которая, наряду с существительными I склонения типа *башкир*, *гренадер*, *глаз*, имеет в родительном падеже множественного числа нулевое окончание.

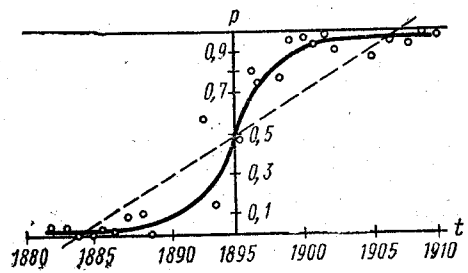


Рис. 15

Статистический ход этого процесса отражен в табл. 1.1, составленный на основании данных, приводимых в работе [17].

Перенесем данные этой таблицы на график (рис. 15), отмечая на оси абсцисс годы (t), а на оси ординат — относительную частоту (f) нулевых форм [столбец (7)]. Полученная последовательность экспериментальных точек показывает резкое возрастание нулевых форм в период между 1886 и 1905 годами. Возникает вопрос: какой из известных нам элементарных функций можно воспользоваться для описания полученной зависимости? Обозначим искомую зависимость в виде

$$f = \varphi(t). \qquad (1.22)$$

Линейная зависимость вида $at + b$ (см. штриховую линию на рис. 15) здесь применена быть не может, поскольку значения такой функции находятся в интервале от $-\infty$ до $+\infty$, в то время как по условию задачи область изменения нашей функции f лежит в интервале между нулем и единицей (относительные частоты не могут быть меньше нуля и больше единицы). Значения же аргумента t могут быть распределены по всему интервалу от $-\infty$ до ∞. Если

Таблица 1.1

Вероятности и частоты употреблений нулевых форм родительного падежа множественного числа у русских единиц измерения типа *вольт* в конце XIX и начале XX в.

t	$t-1895$	$x=\frac{t-1895}{3}$	arctg x	$\frac{1}{\pi}$ arctg x	p	$f=\frac{F}{100}$	$p-f$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1881	—14	—4,66	—1,35	—0,43	0,07	0,06	0,01
1882	—13	—4,33	—1,34	—0,43	0,07	0,06	0,01
1883	—12	—4,00	—1,32	—0,41	0,09	0,01	0,08
1884	—11	—3,66	—1,29	—0,40	0,10	0,00	0,10
1885	—10	—3,33	—1,27	—0,40	0,10	0,01	0,09
1886	—9	—3,00	—1,24	—0,39	0,11	0,01	0,10
1887	—8	—2,66	—1,20	—0,38	0,12	0,03	0,09
1888	—7	—2,33	—1,16	—0,37	0,13	0,09	0,04
1889	—6	—2,00	—1,10	—0,35	0,15	0,11	0,04
1890	—5	—1,66	—1,02	—0,32	0,18	0,00	0,18
1891	—4	—1,33	—0,92	—0,29	0,21	0,57	—0,36
1892	—3	—1	—0,78	—0,25	0,25	0,58	—0,33
1893	—2	—0,66	—0,58	—0,18	0,32	0,59	—0,27
1894	—1	—0,33	—0,31	—0,10	0,40	0,16	0,24
1895	0	0,00	0,00	0,00	0,50	0,47	0,03
1896	1	0,33	0,31	0,10	0,60	0,76	—0,16
1897	2	0,66	0,58	0,18	0,68	0,68	0,00
1898	3	1,00	0,78	0,25	0,75	0,72	0,03
1899	4	1,33	0,92	0,29	0,78	0,96	—0,18
1900	5	1,66	1,02	0,32	0,82	0,98	—0,16
1901	6	2,00	1,10	0,35	0,85	0,95	—0,10
1902	7	2,33	1,16	0,37	0,87	0,98	—0,11
1903	8	2,66	1,20	0,38	0,88	0,90	—0,02
1904	9	3,00	1,24	0,39	0,89	0,52	0,37
1905	10	3,33	1,27	0,40	0,90	0,89	0,01
1906	11	3,66	1,29	0,41	0,91	0,97	—0,06
1907	12	4,00	1,32	0,42	0,92	0,92	—0,00
1908	13	4,33	1,34	0,43	0,93	0,99	—0,06
1909	14	4,66	1,35	0,43	0,93	1,00	—0,07
1910	15	5	1,55	0,44	0,94	0,99	0,05

провести эмпирическую кривую через отмеченные точки, то ее левая ветвь асимптотически приближается к оси Ot, а правая — к прямой $p = 1$. Середина кривой показывает резкий подъем вверх. Указанным условиям лучше всего соответствует график обратной тригонометрической функции (1.20). Однако рассматриваемая эмпирическая кривая имеет ряд особенностей по сравнению с теоретической кривой (см. рис. 14). Во-первых, в качестве оси симметрии наша кривая имеет вертикальную прямую, проходящую не через точку $t = 0$, а через $t = 1895$. Во-вторых, интервал, в котором изменяется значение функции f, равен не π, а единице. В-третьих, кривая поднята над осью Ot на половину интервала, в котором заключены значения функции. Все эти особенности кривой могут быть учтены с помощью коэффициентов, которые следует ввести в выражение (1.20). Так, чтобы отразить сдвиг оси симметрии вправо на 1895 ед., мы должны от аргумента t отнять 1895; в этом случае имеем

t — 1895. В целях более компактного и наглядного построения графика величину t — 1895 нужно уменьшить в три раза. Следовательно, аргументом будет служить величина $x = (t - 1895)/3$. Чтобы учесть различия в размерах интервала, в котором происходит изменение значений функции f, выражение arctg $(t - 1895)/3$ следует умножить на поправочный коэффициент $1/\pi$. Наконец, чтобы поднять интервал значений функций f над осью Ot, к выражению $(1/\pi)$ arctg $(t - 1895)/3$ нужно прибавить 0,5, что равно половине указанного интервала. В итоге получаем равенство

$$p = \frac{1}{\pi} \operatorname{arctg}\left(\frac{t-1895}{3}\right) + 0{,}5, \qquad (1.23)$$

которое является аналитическим выражением, характеризующим рост употребления нулевых форм родительного падежа множественного числа у существительных типа *вольт*, *рентген* в русских научно-технических текстах конца прошлого и начала нынешнего века.

Кривая, описывающая отступление регулярных форм родительного падежа множественного числа тех же существительных (рис. 16), является зеркальным отображением только что проанализированной кривой. Нетрудно поэтому прийти к заключению, что аналитическое выражение зависимости между временем t и относительной частотой употребления регулярных форм

$$f' = \frac{100 - F}{100} = q \qquad (1.24)$$

имеет в этом случае вид

$$q = \frac{1}{\pi} \operatorname{arcctg}\left(\frac{t-1895}{3}\right) + 0{,}5. \qquad (1.25)$$

Правильность этого вывода предлагается проверить читателю.

В только что приведенном примере использование математических методов имеет больше иллюстративный, чем технологический характер.

Однако в некоторых лингвистических задачах моделирование диахронического процесса с помощью элементарных функций служит средством восстановления не засвидетельствованных в памятниках и диалектах этапов исследуемого процесса. Обратимся в этой связи к истории формирования и развития определенного артикля в старофранцузском языке.

2. Формирование определенного артикля во французском языке. Известно, что формирование французского определенного артикля, как и других новых аналитических категорий, осуществлялось в протороманскую эпоху (V—VIII в.) в народно-разговорной речи, которая почти не отражена в дошедших до нас позднелатинских памятниках [27]; [37].

Старофранцузские памятники IX—XIII в. фиксируют уже завершение процесса формирования определенного артикля и установление современных норм его употребления при счисляемых именах существительных. Распространение определенного артикля на несчисляемые (абстрактные, вещественные, Singularia tantum) существительные, а также на некоторые разряды имен собственных происходит в более поздние периоды.

Восстановить начальный период формирования артикля непосредственно по народно-латинским текстам V—VIII в. невозможно, поскольку эти памятники лишь косвенно отражают разговорную речь. Кроме того, если говорить непосредственно о формировании норм употребления артикля, то здесь очень трудно разграничить старые местоименные и новые членные употребления форм латинского детерминатива ille, к которым восходят формы французского определенного артикля. Поэтому приходится использовать различные приемы реконструкции, исходя при этом из старофранцузского материала, где артикль имеет не только четко выраженную форму (синкопированные производные от ille), но также достаточно ясно очерченные нормы употребления и грамматическое значение.

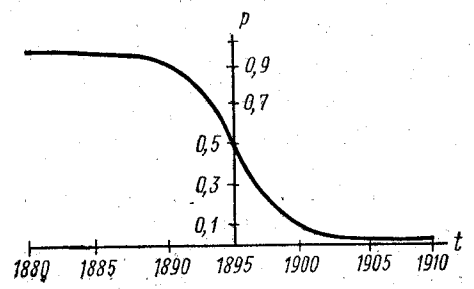


Рис. 16

В этом смысле определенный интерес представляет экстраполяционная реконструкция текстовой статистики употребления артикля в те периоды его истории, которые не имеют памятников разговорной романской речи. С этой целью была определена относительная частота, или частость (f) употребления артикля в тех отрывках из памятников, в которых дана прямая речь героев (предполагается, что прямая речь лучше всего отражает нормы народно-разговорного языка). Однако этого материала оказалось недостаточно. Первым памятником, включающим прямую речь героев, оказалось «Житие св. Алексиса», написанное около середины XI в., и только в этом памятнике частость форм определенного артикля при счисляемых существительных заметно отличается от тех относительных частот его употребления, которые дают более поздние памятники, отражающие завершение интересующего нас процесса (см. табл. 1.2). Иными словами, памятники не дают статистических сведений достаточных для экстраполяционной реконструкции исследуемого процесса. Поэтому приходится обратиться к данным старофранцузской топонимики.

Латинские хартии Галлии VIII—X в. содержат уже большое число топонимов, отражающих не только фонетические, но и грамматические черты живой народно-разговорной романской речи. Немало в этих документах таких названий, которые включают романские артикли (le, la, los): ср. Lagarda — начало IX в., los Grineurios — 900 г. Вообще ранние романские топонимы могли возник-

нуть только на почве устной народно-разговорной речи, поскольку литературная или деловая письменная традиция в романском языке в этот период практически отсутствовала. Отдельные глоссы и предположительно существовавшие памятники типа «Страсбургских клятв» не могли образовать особого функционального стиля.

Известно, что средневековые письменные памятники с опозданием примерно на 100 лет фиксируют явления, сформировавшиеся и укрепившиеся в устной речи. Предположим, что эту хронологическую поправку можно принять и для топонимики. Чтобы проверить все эти предположения применительно к нашей статистической задаче, сопоставим частоты появления артикля в топонимах определенного периода с относительными частотами употребления артикля в разговорных фрагментах памятников, написанных при-

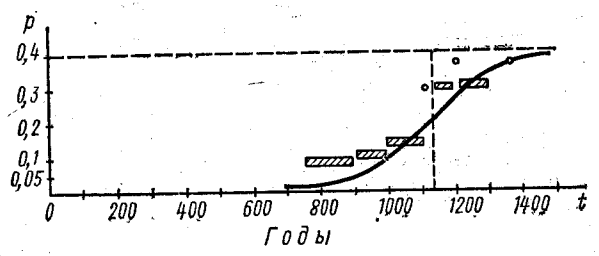


Рис. 17

мерно на сто лет раньше. Если частости артикля в топонимах XII в. соответствовали бы его частостям в разговорных фрагментах XI в., а топонимические частости XIII в. были бы близки к разговорным относительным частотам XII в. и т. д., то это означало бы, что наши хронологические и стилистические поправки взяты правильно. Это в свою очередь дало бы нам возможность использовать статистику артикля по топонимам VIII—XI в., для которых памятники романской разговорной речи отсутствуют. Если же сопоставленные частости давали бы заметные расхождения, то это свидетельствовало бы о том, что наши стилистические и хронологические предположения неверны, и мы не имеем права использовать топонимическую статистику для экстраполяции начального этапа формирования артикля.

Данные табл. 1.2, заимствованной из работы [32 в, с. 370—374], показывают достаточную близость в частостях употребления артикля в соответствующих выборках топонимов и разговорных фрагментах, что дает нам право использовать статистику топонимов для VIII—X веков.

Нанесем на график наши табличные данные и подберем кривую, отражающую динамику роста употребительности артикля в протороманский и старофранцузский периоды (рис. 17). Затем, применяя методику, использованную при моделировании процесса образования нового лексико-морфологического разряда внутри русских су-

Таблица 1.2

Динамика развития употребительности определенного артикля в протороманский и старофранцузский период

Века	Памятники	Учтено существительных		Частость употребления определенного артикля (f)
		Всего	С определенным артиклем	
VIII—IX	Топонимы IX—X в.	245	24	0,096
X	Топонимы XI в.	273	30	0,106
XI	Топонимы XII в.	526	90	0,176
	«Житие св. Алексиса»	232	38	0,164
XII	«Песня о Роланде» (около 1100 г.)	200	63	0,315
	Топонимы XIII в.	991	323	0,326
	«Рауль де Камбрэ» (конец XII века)	157	44	0,280
XIII	«Окассен и Николетта» (начало XIII в.)	200	75	0,375
	Топонимы XIII в.	542	216	0,399
XIV	«Бодуэн де Себур», Г. де Машо			
	«Взятие Александрии», Ж. Фруассар «Debat dou cheval et dou levrier»	200	74	0,370

ществительных (см. п. 1 настоящего параграфа), найдем аналитическое выражение нашей кривой, которое имеет вид

$$p = \frac{0,4}{\pi} \operatorname{arctg}\left(\frac{t-1125}{100}\right) + 0,2. \qquad (1.26)$$

Условные обозначения здесь те же, что и в зависимости (1.23).

Математическое описание общей динамики развития определенного артикля имеет здесь не просто иллюстративный характер. Оно может помочь при описании деталей формирования этой категории. Так, например, высказывались предположения о том, что регулярное употребление артикля (или протоартикля) установилось в первую очередь при тех именах существительных, значение которых предусматривало устойчивую ситуативную соотнесенность обозначаемого ими предмета с некоторым количеством однородных предметов, а также при тех существительных, которые обозначают предметы или понятия, находящиеся в центре внимания собеседников [32 в, с. 373]. Предварительное статистическое обследование народно-латинских памятников VI—IX в. (Григорий Турский, путешествие Этерии, книга истории Франков, Фредегарий Схоластик, Салическая правда, Клунийские хартии) показывают, что существительные, которые обладают только что указанными признаками и имеют при себе детерминативы, восходящие к ille или ipse, составляют от 3% до 5% всех существительных текста. Это хорошо согласуется с реконструированной нами с помощью математической экстраполяции относительной частотой употребления протороманского артикля в эпоху раннего средневековья (рис. 17).

3. Диахронический скачок и сдвиги в структуре языка. Только что построенные с помощью обратных тригонометрических функций модели можно использовать при описании структурных сдвигов не только в области лексики и морфологии, но также и в других сферах языка — фонологии, синтаксисе, стилистике. Такие сдвиги обнаруживаются либо в появлении новых элементов, либо в отступлении старых лингвистических единиц, причем структурные изменения языка происходят скачкообразно, что особенно отчетливо прослеживается в диахронической статистике текстов. Действительно, образование новой структурной лингвистической единицы сопровождается обычно быстрым ростом употребительности тех форм (звуков, букв, морфем, суффиксов, слов, конструкций), которые служат средством выражения новой фонемы, лексической или грамматической категории. При этом в сравнительно короткий промежуток времени происходит переход от старого, более низкого уровня употребительности соответствующих языковых форм к новому, заметно более высокому уровню использования этих форм. Этот диахронический скачок хорошо моделируется с помощью обратной тригонометрической функции (1.20). Напротив, исчезновение той или иной единицы из системы языка сопровождается обратным перепадом частот, который описывается обычно с помощью функции (1.21).

§ 8. Моделирование информационного построения речи

1. Измерение синтактической информации в речи. Количественные оценки построения речи опираются на квантитативный анализ распределения в ней синтактической информации. Синтактическая информация характеризует структурную организацию текста, через нее можно количественно оценить и смысловую информацию, содержащуюся в словах, словосочетаниях и предложениях текста [26]. Как синтактическая, так и семантическая информация измеряется в двоичных единицах (битах) [41, с. 70]. Более строгое определение понятия «синтактическая информация» дано ниже (см. гл. 5, § 5, п. 5), там же описаны приемы ее вычисления. Сейчас наша задача состоит в том, чтобы найти некоторую математическую модель, аппроксимирующую распределение этой информации в слове и связном тексте.

2. Распределение информации в слове. Как показывают данные табл. 1.3, описывающей распределение синтактической информации I в слове, начальные буквы русского письменного слова несут заметно больше информации, чем буквы, находящиеся в его середине и на конце. Это предположение подтверждается данными экспериментов, проведенных на материале ряда индоевропейских и тюркских языков [23, с. 80—89]. Какой функцией можно было бы воспользоваться для моделирования этого убывания информации при движении по слову слева направо? Если отвлечься от небольших «пиков» и «провалов» информации в середине и конце слова, которые условно можно отнести за счет статистического разброса, а также если считать, что теоретически длина слова в большинстве языков ничем

не ограничена (подробнее см. гл. 4, § 2, п. 1), то распределение информации в слове можно представить в виде пологой экспоненты, асимптотически приближающейся к оси абсцисс (рис. 18). Аналитическое выражение такой кривой имеет вид

$$I_n = I_0 e^{-sn}. \qquad (1.27)$$

Строгое доказательство этого утверждения приведено в гл. 2, § 4. Выражение (1.27) есть усложненный вариант экспоненциальной функции (1.10), где в качестве аргумента выступает номер буквы n, а в качестве функции I_n — количество информации, которую несет буква n. Формула (1.27) содержит два параметра I_0 и s, имеющие лингвистический смысл. Постоянная I_0 (ее называют обычно *информацией алфавита*) показывает максимальную величину информации, которую несла бы буква языка, использующего алфавит S, если в системе и норме этого языка не было заложено ограничений на сочетаемость букв и вероятности их употребления. Параметр s показывает темп нарастания ограничений, накладываемых системой и нормой языка на неопределенность выбора n-й буквы слова при условии, что цепочка букв, находящаяся слева, уже известна.

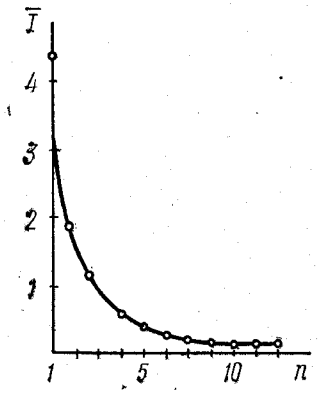


Рис. 18

Таблица 1.3

Распределение синтактической информации в обобщающих схемах текстового русского слова и текста* по $\bar{I}$

Буквы	1	2	3	4	5	6	7	8
Значения $\bar{I}$ в слове	3,45	1,90	1,15	0,55	0,35	0,28	0,25	0,20
Значения $\bar{I}$ в тексте	4,22	2,85	3,16	2,45	2,28	2,54	2,12	2,22
Буквы	9	10	11	12	15	20	25	30
Значения $\bar{I}$ в слове	0,17	0,20	0,15	0,10	—	—	—	—
Значения $\bar{I}$ в тексте	2,05	1,65	2,35	2,07	1,86	1,37	1,67	1,05

* Ниже будет показано (см. гл. 5, § 5, п. 4), что существующие методы позволяют определять лишь верхнюю ($\bar{I}$) и нижнюю ($\underline{I}$) границы интервала, в котором находится истинное значение информации I. Поэтому при решении информационно-лингвистических задач используются либо верхние, либо нижние оценки информации.

3. Распределение информации в тексте. Приведенные выше рассуждения можно распространить и на связный текст (см. табл.1.3). Тогда математической моделью распределения информации станет показательная функция

$$I_n = (I_0 - I_\infty)\, e^{-sn} + I_\infty, \qquad (1.28)$$

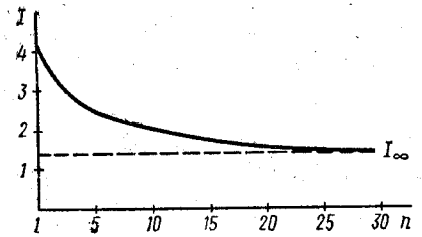


Рис. 19

отличающаяся от (1.27) тем, что в нее введен новый параметр I_∞, указывающий на тот предельный уровень, к которому асимптотически приближается средняя величина информации на букву при бесконечном увеличении длины текста (рис. 19). Величина I_∞ указывает на то, что каждый текст в отличие от слова может быть продолжен и дальше и это продолжение несет в себе некоторую информацию. Более строгое обоснование для параметра I_∞ приведено в гл. 2, §4 вместе с доказательством равенства (1.28). Значения I_∞ для разных языков показаны* в табл. 1.4.

Таблица 1.4

Значения $\overline{I}_\infty$ и $\underline{I}_\infty$ для некоторых индоевропейских и тюркских языков

	Язык	$\overline{I}_\infty$	$\underline{I}_\infty$		Язык	$\overline{I}_\infty$	$\underline{I}_\infty$
1	Русский	1,37	0,82	5	Французский . .	1,38	0,79
2	Польский	1,28	0,76	6	Румынский . . .	1,34	0,72
3	Английский . . .	1,35	0,74	7	Казахский	1,51	0,82
4	Немецкий	1,36	0,71				

4. Распределение контекстной обусловленности. Выше уже говорилось, что I_0 — максимальная синтактическая информация, которая может быть получена от лингвистического элемента, входящего в алфавит (парадигму) S, при условии, что вероятностно-комбинаторные ограничения здесь не учитываются. Поскольку I_n — реальная информация, извлекаемая из лингвистического элемента в участке текста n, то разность K_n этих информаций будем называть *контекстной обусловленностью*:

$$K_n = I_0 - I_n. \qquad (1.29)$$

Величину K_n можно рассматривать как меру тех структурных, нормативных ограничений, которые накладываются на букву или любой другой лингвистический элемент, находящийся в n-й позиции текста.

Подставляя в зависимость (1.29) вместо I_n выражение (1.28) и производя необходимые преобразования, приходим к формуле

$$K_n = (I_0 - I_\infty)\,(1 - e^{-sn}), \qquad (1.30)$$

выражающей рост контекстной обусловленности как функцию числа n. Значения параметров здесь те же, что и в соотношениях (1.27) и (1.28). На рис. 20 показаны кривые, характеризующие нарастание контекстных связей в русских беллетристических и деловых текстах; в табл. 1.5 приведены значения коэффициента s, характеризующего темп роста этих связей в тексте. Вполне естественно, что в деловых текстах, использующих стандартизованную терминологию, фразеологию и синтаксис, контекстные связи растут быстрее, чем в беллетристических текстах, пользующихся более разнообразной лексикой, фразеологией и синтаксической вариативностью.

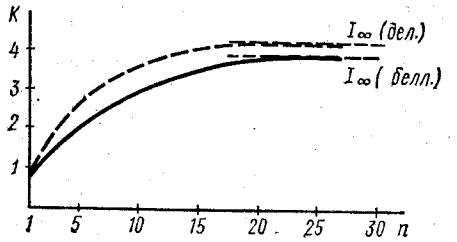


Рис. 20

Таблица 1.5

Предельная информация и контекстный коэффициент для русских текстов

Разновидности русской речи	На верхней границе информации	
	I_∞	s
Беллетристика	1,19	0,21
Деловые тексты	0,83	0,24

Пределом, к которому стремится экспонента контекстной обусловленности, служит *предельная контекстная обусловленность* $K_\infty = I_0 - I_\infty$.

§ 9. Моделирование периодичности речи

1. Гармоническая структура гласных. Наблюдения над распределением информации в тексте и слове (см. рис. 18 и 19), а также изучение звуковой субстанции позволяют обнаружить в этих явлениях некоторую периодичность. Лучше всего эта периодичность прослеживается в построении гласных звуков.

Однако гласный звук следует рассматривать не как элементарную, а как сложную периодическую функцию. Чтобы дать математическое описание этой функции, нужно представить ее в виде суммы простейших тригонометрических функций.

* См. сноску к табл. 1.3 на стр. 39.

Рассмотрим это представление на примере искусственного сложного звука, который можно считать упрощенным аналогом гласного. Чтобы получить искусственный гласный, включим два камертона — первый с частотой F, равной 200 колебаниям в секунду (герц, сокращенно — Гц), и второй — с частотой $F = 600$ Гц. Тогда графическое изображение функции, которая характеризует сложный звук, производимый обоими камертонами вместе, будет представлять собой *сложную гармоническую кривую* C, показанную на рис. 21. Такая кривая является результатом сложения двух синусоид, аргументом которых служит время (эти синусоиды называют-

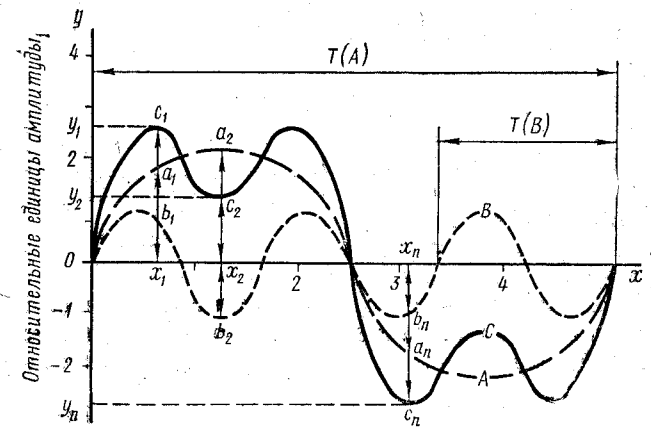


Рис. 21

ся *гармониками**). Синусоида A характеризует звуковые колебания, производимые первым камертоном, а синусоида B — вторым. При этом каждая точка C нашей сложной кривой имеет ординату y_C, представляющую собой сумму ординат y_A и y_B, или

$$y_C = f(x) + \Phi(x). \qquad (1.31)$$

Мы использовали различные обозначения функций f и Φ, поскольку синусоиды A и B имеют разные по абсолютной величине периоды колебаний и разные амплитуды.

Аргументы функций $f(x)$ и $\Phi(x)$ измеряются в единицах времени (в мс). Переведем значения аргумента в радианную (или угловую) меру. Так как длина периода гармоники равна T, то каждое значение аргумента x составляет некоторую долю этого периода, равную отношению x/T. Но длина периода T, взятая в радианной мере, равна 2π. Отсюда следует, что, умножив отношение x/T на 2π, мы получим значение аргумента в виде $2\pi x/T$, где частное $2\pi/T = \omega$ есть коэффициент перехода от линейного (временного) измерения аргумента к его радианной мере.

Найдем значения коэффициентов перехода в рассматриваемом примере. Так как $T_A = 1/F_A$, $T_B = 1/F_B$, $F_A = 200$ Гц, $F_B = 600$ Гц, то $T_B = T_A/3$. Следовательно, полагая $T_A = T$, получим:

$$\omega_A = 2\pi/T_A = 2\pi/T, \quad \omega_B = 6\pi/T_A = 6\pi/T.$$

Таким образом, для гармоник A и B имеем

$$\sin \omega_A x = \sin (2\pi x/T), \qquad (1.32)$$

$$\sin \omega_B x = \sin (6\pi x/T). \qquad (1.33)$$

Однако эти величины еще не являются значениями y_A и y_B, ведь амплитуды r_A и r_B обеих синусоид не равны между собой.

Амплитуды гармонических кривых измеряются и в единицах длины, и в единицах интенсивности (для звуковой волны). Чтобы не привязывать наши рассуждения к какой-либо узкой среде приложения периодических функций, мы для сопоставительного измерения амплитуд воспользуемся относительными единицами, при этом $r_A = 2{,}3$ отн. ед., а $r_B = 1$ отн. ед. Чтобы получить значения интересующих функций, нужно умножить выражения (1.32) и (1.33) соответственно на коэффициенты r_A и r_B. При этом

$$y_A = f(x) = r_A \sin \omega_A x, \quad y_B = \Phi(x) = r_B \sin \omega_B x,$$

откуда согласно (1.31) получаем

$$y_C = r_A \sin \omega_A x + r_B \sin \omega_B x. \qquad (1.34)$$

Взяв численные значения параметров r_A и r_B, получим выражение для ординаты любой точки сложной гармонической кривой:

$$y_C = 2{,}3 \sin (2\pi x/T) + \sin (6\pi x/T).$$

Разложение сложной гармонической кривой на составляющие ее синусоиды в том виде, как оно представлено на рис. 21, оказывается сложным и не очень наглядным даже при наложении двух гармоник. Если же сложная кривая является результатом взаимодействия многих синусоид, то графическое изображение их взаимодействия может оказаться совершенно лишенным наглядности. Поэтому обычно используется упрощенное изображение структуры (или, как говорят, *спектра*) сложной кривой. При этом на оси абсцисс откладывается частота колебания, а на оси ординат — величина амплитуды в единицах длины или интенсивности (децибелах — сокращенно дБ) или, наконец, в относительных единицах. Эти схемы называются *спектрограммами*. На рис. 22 показана спектрограмма сложного звука, полученного от наших двух камертонов: ле-

* В исследованиях по акустике речи гармониками называются обычно только дополнительные синусоиды (частоты), накладывающиеся на так называемую несущую или основную частоту (синусоиду) голоса. Основную частоту и дополнительные частоты можно представить также в виде обратной пропорциональной зависимости $F = 1/T$, аргумент T которой является периодом колебания, измеряемым в единицах времени.

вый столбик указывает сильную амплитуду гармоники A, правый — слабую амплитуду гармоники B. Что касается реальных гласных, то характеризующие эти звуки кривые также могут быть представлены как суммы многочисленных гармоник.

Разложение сложной кривой можно произвести либо вручную с помощью приемов гармонического анализа, о чем мы будем говорить ниже, либо автоматически — путем использования специальных приборов, выделяющих составляющие звук гармоники (такими приборами являются, в частности, спектрометр и спектрограф).

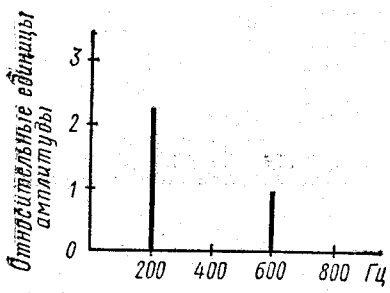


Рис. 22

С помощью автоматического разложения кривой речевого сигнала обычно можно обнаружить несколько десятков гармоник: одни из них имеют более сильные, другие — более слабые амплитуды. Усиленные гармоники расположены компактно в нескольких областях звукового спектра, которые называются *формантными областями* (*формантами*), или *областями концентрации энергии* в спектре звука речи. Эти форманты обозначаются обычно символами F_0, F_1, F_2, F_3 и т. д.

При акустическом анализе речи гармоники со слабыми амплитудами не учитываются. Исследователь сосредоточивает свое внимание на расположении и соотношении формантных областей, давая им сначала физиологическую, а затем лингвистическую интерпретацию*. В качестве примера рассмотрим спектрограммы трех видов эстонского гласного [õ] — краткого [õ], долгого [õ:] и сверхдолгого [õ: :] (рис. 23, *а*, *б* и *в*). На фотографии видно, как гармоники (аппарат отметил их в виде столбиков) группируются в форманты (F). Каждая форманта имеет одну или несколько наиболее сильных гармоник (f_m). Формантные области и усиленные гармоники каждой из этих формант приведены в табл. 1.6, заимствованной из работы** [54].

Получаемые с помощью гармонического анализа и автоматической спектрографии данные о формантной структуре звуков имеют

* Форманты также описываются сложными кривыми, которые могут быть представлены в виде суммы гармоник.

** Автор этой работы Г. Лийв следующим образом интерпретирует приведенные экспериментальные данные.

«Относительно спектрального состава вариантов гласного [õ] разных степеней долготы следует отметить, что в связи с увеличением степени долготы первая форманта понижается (вероятно, это акустический коррелат сужения артикуляции); вторая и более высокие форманты также в общем понижаются, причем более значительный сдвиг в сторону более низких частот вместе с понижением относительного уровня высоких формант выступает при [õ ::] третьей степени долготы (вероятно, это акустический коррелат большей велярности артикуляции)» [54, с. 97].

также большое практическое значение. Они используются при проектировании автоматов, воспринимающих и воспроизводящих человеческую речь.

2. Разложение сложной периодической кривой в ряд. Для выявления периодичности в информационной структуре текста необходим более сложный математический аппарат, чем тот, который использовался при моделировании сложного звука. Рассмотрим в этой связи выражение (1.34). Так как $\sin x = \cos(x - \pi/2)$, то указанное выражение можно переписать в виде

$$y = r_A \cos(\omega_A x - \pi/2) + r_B \cos(\omega_B x - \pi/2).$$

Отсюда следует, что каждая периодическая кривая может быть сдвинута по отношению к ее началу на некоторую величину φ, которая называется *сдвигом фазы* (φ может быть и меньше и больше, чем $\pi/2$). Чтобы учесть этот сдвиг, величину φ надо ввести под знак тригонометрической функции; при этом получится равенство вида

$$y = r \sin(\omega x + \varphi)$$

или

$$y = r \cos(\omega x + \varphi).$$

Колебания функции могут происходить не обязательно относительно оси абсцисс, как это мы наблюдали до сих пор. Периодическая функция может быть сдвинута вверх или вниз по отношению к оси Ox. Чтобы отразить этот сдвиг, нужно ввести в каждое из равенств постоянный член r_0, соответствующий величине ординаты, на которую поднята или опущена ось, вокруг которой происходят колебания нашей периодической кривой. В результате получается равенство вида

$$y = r_0 + r \sin(\omega x + \varphi)$$

или

$$y = r_0 + r \cos(\omega x + \varphi).$$

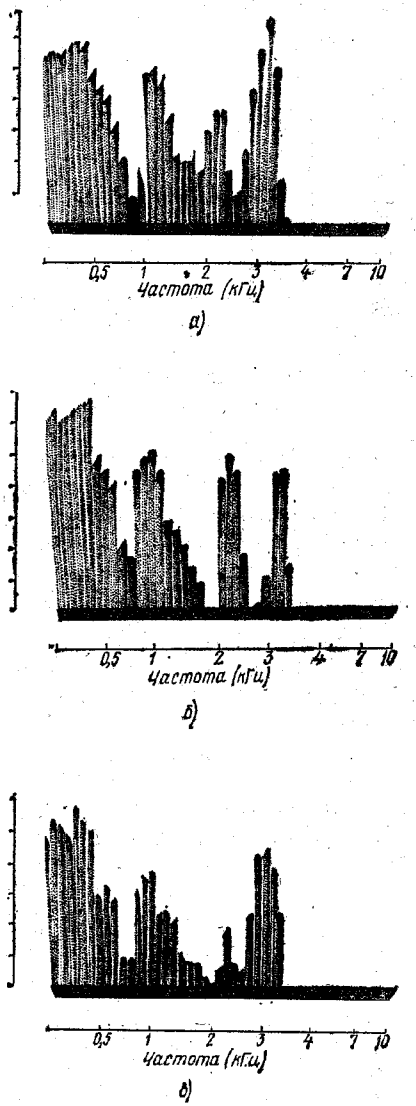


Рис. 23

Таблица 1.6

Формантные области и их усиленные гармоники для трех эстонских [õ]

Форманты	[õ]	[õ:]	[õ::]
F_1	до 525 (Гц)	до 525 (Гц)	до 525 (Гц)
fm_1	450	450—525	375
F_2	1150—1350	975—1250	975—1150
fm_2	1250	1150	1150
F_3	1950—2250	2100—2400	около 2250
fm_3	2100—2250	2250	—
F_4	2850—3300	3150—3300	2850—3150
fm_4	3150	3300	3000

Теперь предположим, что имеется некоторая сложная периодическая функция y (сложная гармоническая кривая), поднятая над осью абсцисс на величину r_0. Тогда мы можем представить эту функцию в виде ряда слагаемых, каждое из которых имеет вид

$$y_k = r_k \sin (2\pi kx/T + \varphi_k).$$

Слагаемые представляют собой синусоидальные гармоники с амплитудой r_k и сдвигом фазы φ_k (рис. 24). Частоты гармоник, из которых составляется периодическая функция y_k, образуют гармоническую последбательность. Это значит, что частоты всех составляющих этой функции кратны основной частоте $1/T$, причем синусоида y_1 с частотой $1/T$ служит первой гармоникой ($k=1$), синусоида y_2 с частотой $2/T$ — второй гармоникой ($k = 2$),... , синусоида y_m с частотой m/T — m-й гармоникой ($k = m$). Величина r_0, выражающая среднее значение функции y и равная ординате, на которую поднята ось сложной периодической кривой и составляющих ее гармоник, рассматривается в качестве нулевой гармоники. При этих условиях наша сложная периодическая кривая может быть представлена в виде суммы, называемой *рядом Фурье*:

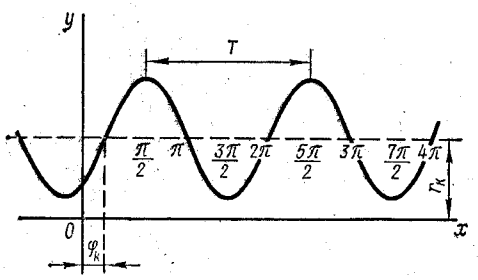


Рис. 24

$$y = r_0 + r_1 \sin\left(\frac{2\pi}{T}x + \varphi_1\right) + r_2 \sin\left(\frac{2\pi}{T}2x + \varphi_2\right) + \ldots$$

$$\ldots + r_m \sin\left(\frac{2\pi}{T}mx + \varphi_m\right),$$

или в сокращенной записи

$$y = r_0 + \sum_{k=1}^{m} r_k \sin (2\pi kx/T + \varphi_k). \qquad (1.35)$$

3. Периодичность в информационной схеме текста. Используем приведенные выше сведения для анализа лингвистического материала. Если побуквенно угадывать не связанные между собой

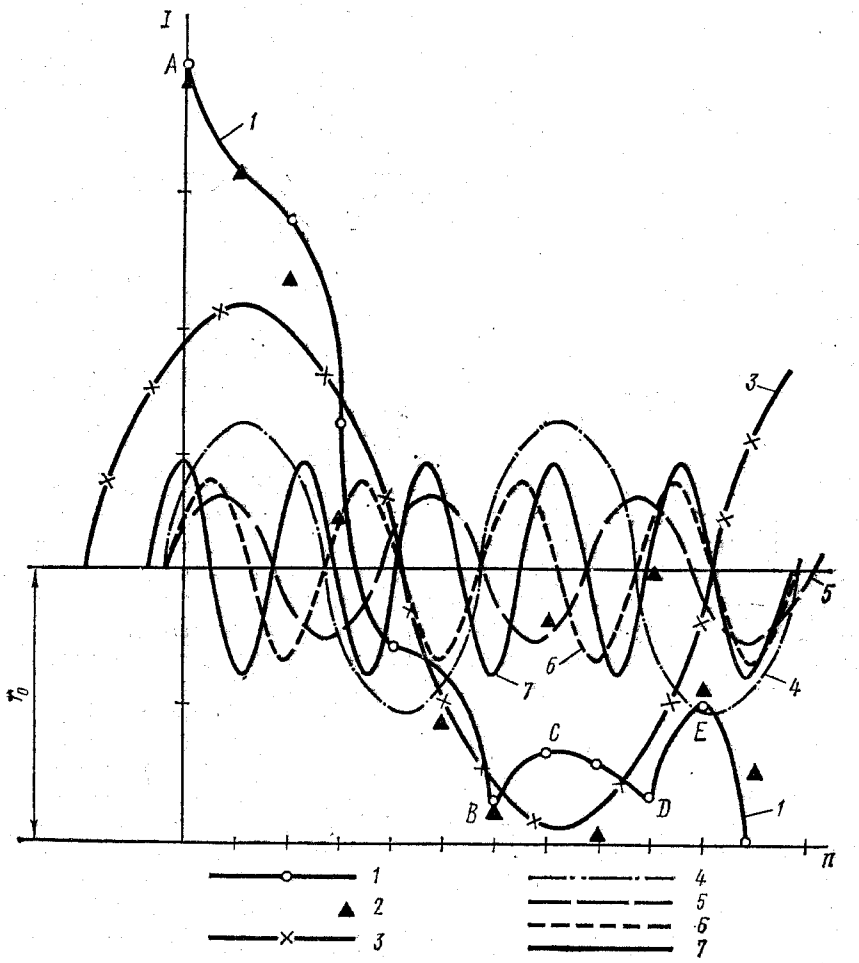


Рис. 25. Разложение сложной гармонической кривой распределения информации на составляющие ее гармоники: *1* — кривая и эмпирические точки нижней границы информации в 12-буквенной словоформе, взятой вне контекста; *2* — теоретические точки; *3* — первая гармоника; *4* — вторая гармоника; *5* — третья гармоника; *6* — четвертая гармоника; *7* — пятая гармоника

по смыслу слова определенной длины, то в результате обработки получаемых данных по методу, описанному в гл. 5, § 5, п. 4, мы получим кривую с повторяющимся распределением информации.

На основании приведенных в работе [23, табл. 20] данных о нижней границе распределения информации в словах разной длины, взятых вне текста, было осуществлено разложение каждой из этих кривых в ряд. Это разложение, производившееся с помощью схем группирования Рунге-Серебренникова [32в, с. 397], удовлетворяет следующему равенству:

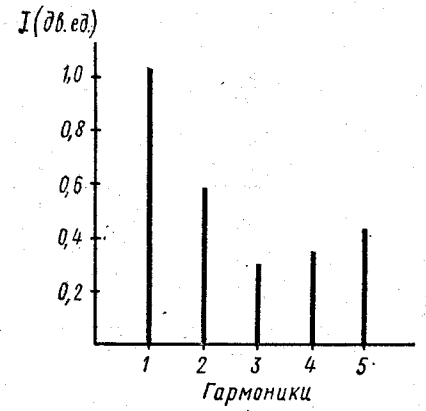


Рис. 26

$$I = r_0 + \sum_{k=1}^{5} r_k \sin(2\pi k x / l + \varphi_k), \qquad (1.36)$$

где I — информация; x — непрерывный аргумент функции, заменяющий дискретную величину $n - 1$; n — номер буквы слова; r_0 — нулевой член ряда Фурье (соответствует среднему арифметическому значению информации на рассматриваемом участке); r_k — амплитуда k-й гармоники ряда Фурье (величина информации k-й гармоники); l — длина слова в буквах, т. е. длина участка, на котором производится разложение; φ_k — сдвиг фазы k-й гармоники в радианах. Результаты разложения кривых, описывающих распределение информации русских слов длиной в 5—12 букв (включая пробелы), приведены в табл. 1.7.

На рис. 25 показано разложение сложной гармонической кривой, характеризующей распределение информации (по нижней границе) в 12 буквенном слове, на составляющие ее гармоники. На рис. 26 показана спектрограмма этого распределения.

Анализируя табл. 1.7, нетрудно заметить, что первые гармоники дают обычно наибольшие амплитуды, которые ослабляются у гармоник более низкого порядка. Однако обращают на себя внимание случаи усиления амплитуд у тех гармоник, период которых находится в интервале между двумя и тремя с половиной буквами (в таблице они выделены жирным шрифтом). Если учесть, что именно в указанном интервале находятся средние длины корневых морфем русских слов* [23, с. 83—84], то это дает нам возможность предположить, что такие усиленные гармоники являются математическими соответствиями корневых морфем русских словоформ.

Гармонический анализ информационных схем слова и текста имеет для языкознания большой теоретический и практический интерес.

* Средняя длина морфемы в русском языке равна 2,5 буквы.

Таблица 1.7

Амплитуды, сдвиги фаз и периоды гармоник при разложении кривых распределения информации русских словоформ, взятых вне контекста (распределение берется по нижней границе информации)

Длина слова в буквах l	Номер гармоники k	Амплитуда k-й гармоники в дв. ед. r_k	Сдвиг фазы в радианах φ_k	Длина периода в буквах l/k
(1)	(2)	(3)	(4)	(5)
5	1	0,84	0,363	5
	2	**0,78**	0,407	**2,5**
	3	0,52	0,840	1,7
	4	0,33	1,134	1,2
	5	0,36	2,021	1
6	1	0,76	0,209	6
	2	**0,66**	0,785	**3**
	3	0,38	1,358	2
	4	0,38	1,085	1,5
	5	0,35	1,972	1,2
7	1	1,11	0,797	7
	2	**0,79**	0,416	**3,5**
	3	**0,38**	0,747	**2,3**
	4	0,32	1,061	1,7
	5	0,41	2,129	1,4
8	1	1,11	0,590	8
	2	0,37	6,251	4
	3	**0,51**	1,045	**2,6**
	4	0,48	0,948	**2,0**
	5	0,43	2,213	**1,6**
9	1	1,08	0,837	9
	2	0,50	0,020	**4,5**
	3	**0,51**	1,250	**3**
	4	**0,66**	1,390	**2,2**
	5	0,40	2,001	**1,8**
10	1	0,92	0,686	10
	2	0,34	5,771	5
	3	**0,60**	0,628	**3,3**
	4	0,16	0,829	2,5
	5	0,38	1,727	2
11	1	0,99	0,925	11
	2	**0,56**	0,465	**5,5**
	3	0,20	0,354	3,4
	4	0,22	1,347	2,5
	5	0,33	2,117	2,2
12	1	1,03	0,980	12
	2	0,57	0,337	6
	3	0,28	0,564	4
	4	0,35	0,605	3
	5	**0,41**	1,791	**2,4**

Во-первых, он еще раз показывает, что текст имеет квантовую информационную структуру, где информационно нагруженные элементы чередуются со слабоинформативными элементами заполнения. Такое «зернистое» строение текста связано, очевидно, с ритмом работы нейронов мозга, периодически накапливающих и отдающих информацию, поступающую к ним от органов чувств или из других участков мозга.

Во-вторых, объективное выделение информационных гармоник слов и текста позволяет поставить на твердую основу морфолого-синтаксическую типологию языков и создать некоторые количественные эталоны. Сопоставляя с этими эталонами информационные гармоники конкретных текстов, лингвисты получают возможность пролить свет на типологию и происхождение таких загадочных языков, каким является, например, кетский язык.

ГЛАВА 2

ГЛОТТОХРОНОЛОГИЯ, ИНФОРМАЦИОННАЯ СХЕМА ТЕКСТА И ИХ МОДЕЛИРОВАНИЕ С ПОМОЩЬЮ АППАРАТА БЕСКОНЕЧНО МАЛЫХ ВЕЛИЧИН И ПРЕДЕЛОВ

§ 1. Понятия бесконечно малой величины и предела в квантитативной лингвистике

В предыдущей главе мы познакомились с математическими приемами, позволяющими представить в виде аналитической схемы и охарактеризовать числом различные лингвистические процессы. При этом мы оперировали двумя основными математическими понятиями: независимой переменной величиной (это понятие позволяет формализовать всякое лингвистическое изменение и эволюцию) и функцией (это понятие дает возможность следить за количественным и качественным изменением лингвистического явления в зависимости от изменения его аргумента). Однако рассмотренные нами элементарные функции представляют собой довольно грубые схемы-эталоны, которых явно недостаточно, чтобы исследовать любые малые, непрерывно текущие и изменяющиеся по своему темпу диахронические процессы, а также неустановившиеся состояния языка и речи. Для описания этих процессов следует применять более гибкий математический аппарат, построенный на понятиях предела и бесконечно малой величины.

1. Бесконечно малая величина. Нам уже известно (см. гл. 1, § 3, п. 2), что переменной называется величина, принимающая различные численные значения в течение некоторого лингвистического процесса.

Переменная величина, которая изменяется таким образом, что в процессе своего изменения становится и в дальнейшем остается меньше сколь угодно малого наперед заданного положительного числа ε, носит название *бесконечно малой*. Кратко это определение записывается в виде неравенства $|\alpha| < \varepsilon$, где α — бесконечно малая величина.

Например, бесконечно малая величина может быть получена как вероятность фонологической системы принять устойчивое положение при однократном случайном изменении одного из дифференциальных признаков этой системы [24]. Указанная вероятность определяется выражением

$$p = 1/2^n. \tag{2.1}$$

Предположим, что мы будем увеличивать число дифференциальных признаков фонологической системы, последовательно придавая величине n значения 0, 1, 2, 3, 4, ...; тогда вероятность p принимает соответственно значения 1, 1/2, 1/4, 1/8, 1/16, ...

Очевидно, какое бы малое положительное число ε мы ни взяли, среди значений p всегда найдется число, меньшее чем ε.

Если в качестве ε выбрать число 1/5000, то при $n = 13$ переменная p примет значение 1/8192, меньшее чем ε; если же за ε взять

1/10000, то при $n = 14$ получим, что $p = 1/16384$, т. е. снова* $p < \varepsilon$.

В дальнейшем для нас будут важны следующие свойства бесконечно малых величин.

1. *Алгебраическая сумма конечного числа бесконечно малых величин есть также бесконечно малая величина*, т. е.

$$|\alpha_1 + \alpha_2 + \ldots + \alpha_n| = \sum_i^n \alpha_i < \varepsilon.$$

В том случае, когда число слагаемых неограниченно возрастает, такая сумма может и не быть бесконечно малой величиной.

2. *Произведение бесконечно малой величины α на ограниченную величину x есть также величина бесконечно малая*, т. е.

$$\alpha x < \varepsilon.$$

Следствие 1. *Произведение бесконечно малой величины α на постоянную величину c также есть величина бесконечно малая*, т. е.

$$\alpha c < \varepsilon.$$

Следствие 2. *Произведение двух или нескольких бесконечно малых величин есть величина бесконечно малая*, т. е.

$$|\alpha_1 \alpha_2 \ldots \alpha_n| = \prod_i^n \alpha_i < \varepsilon.$$

2. Предел. Как было только что указано, вероятность фонологической системы принять устойчивое положение при однократном случайном изменении одного из n дифференциальных признаков определяется выражением $p = 1/2^n$.

Если аргументу n последовательно придавать все возрастающие целочисленные значения, то величина p будет неограниченно уменьшаться, приближаясь к нулю. Однако, последовательно уменьшаясь, эта величина всегда остается больше нуля. Следовательно, всегда найдется какая-то бесконечно малая величина, которая по абсолютной величине представляет собой разность между значением переменной и величиной предела ее изменения — в данном случае нулем. Все эти рассуждения приводят нас к понятию предела переменной величины, которое формулируется следующим образом: постоянное число a называется *пределом переменной величины u*, если абсолютная величина их разности есть величина бесконечно малая, т. е. $|u - a| = \alpha$.

Символически это записывается в виде выражения $\lim u = a$, которое читается: «лимит u равен a».

Всякая переменная величина, имеющая предел, является ограниченной — такова, например, величина p в равенстве (2.1).

Определив понятие предела переменной величины, заметим, что обычно переменная величина выступает в виде функции. Так, в соотношении (2.1) переменная величина p является, по существу, функцией, зависящей от n, т. е. $p = f(n)$. Предел такой переменной величины правильнее записать в виде

$$\lim_{n \to \infty} p = 0,$$

указывая в этой записи характер изменения аргумента.

Понятие предела функции можно сформулировать следующим образом. Если любая последовательность значений аргумента x стремится к некоторому числу a, т. е. $\lim_{n \to \infty} x_n = a$, и последовательность значений функции при этом стремится к числу b, то число b называется *пределом функции $y = f(x)$* при условии $x \to a$. Это записывается так:

$$\lim_{x \to a} y = b.$$

Рассмотрим понятие предела функции на примере роста употребительности нулевых форм родительного падежа множественного числа у существительных, обозначающих единицы измерения (*вольт — вольтов, рентген — рентгенов*) — см. гл. 1, § 7, п. 1. Указанный процесс описывается зависимостью (1.23):

$$p = \frac{1}{\pi} \operatorname{arctg}\left(\frac{t - 1895}{3}\right) + 0{,}5,$$

или в общем виде $p = f(t)$, где p — вероятность появления нулевых форм, а t — время.

Возьмем такую последовательность значений аргумента: 1905, 1907, 1908, 1909, 1910, ..., 1920, ..., 1950, ...

Вычислим по формуле (1.23) соответствующие каждому из этих значений t значения функции $f(t)$:

0,906, 0,922, 0,927, 0,933, 0,937, ..., 0,962, ..., 0,983, ...

Очевидно, что по мере приближения t к ∞ рассматриваемая функция стремится к единице.

Приведем теперь основные теоремы о пределах, которые понадобятся нам впоследствии при решении лингвистических задач.

Теорема 1. *Предел алгебраической суммы нескольких переменных величин равен алгебраической сумме пределов этих переменных величин*, т. е.

$$\lim (x + y + \ldots + z) = \lim x + \lim y + \ldots + \lim z.$$

---

* В выражении

$$I_0 = \log_2 S,$$

где $I_0(S)$ — информация алфавита, состоящего из S букв, переменная величина S принимает последовательно значения натурального ряда чисел. Если этот процесс возрастания происходит неограниченно, то какое бы большое положительное число M мы ни взяли, среди значений I_0 всегда найдется число, большее чем M, т. е. будет выполняться неравенство $|I_0| > M$. Такая переменная величина называется *бесконечно большой*.

Между бесконечно большой и бесконечно малой величинами существует связь, которая выражается следующим образом: если A — величина бесконечно большая, то обратная величина $1/A$ есть бесконечно малая; наоборот, если α — бесконечно малая величина (т. е. $\alpha \to 0$), то $1/\alpha$ — бесконечно большая ($1/\alpha \to \infty$).

Теорема 2. *Предел произведения нескольких переменных величин равен произведению пределов этих переменных величин*, т. е.

$$\lim (x \cdot y \ldots z) = \lim x \cdot \lim y \ldots \lim z.$$

Из последней теоремы вытекают три следствия.

Следствие 1. *Предел произведения постоянной на переменную величину равен произведению постоянной на предел переменной величины.* Иными словами, *постоянный множитель можно выносить за знак предела*:

$$\lim ax = \lim a \cdot \lim x = a \lim x.$$

Следствие 2. *Предел целой положительной степени переменной величины, имеющей предел, равен той же степени предела этой переменной величины*:

$$\lim x^n = (\lim x)^n. \qquad (2.2)$$

Следствие 3. *Предел корня m-й степени из переменной величины, имеющей предел, равен корню той же степени из предела этой переменной величины*, т. е.

$$\lim \sqrt[m]{x} = \sqrt[m]{\lim x}. \qquad (2.3)$$

Так как равенство (2.3) можно представить в виде

$$\lim x^{1/m} = (\lim x)^{1/m},$$

то следствие 2 справедливо и для дробных положительных степеней.

Теорема 3. *Предел частного двух переменных величин, имеющих пределы, равен частному от деления этих пределов (при условии, что предел делителя не равен нулю)*, т. е.

$$\lim \frac{x}{y} = \frac{\lim x}{\lim y}.$$

Из теоремы 3 и следствия 2 вытекает, что

$$\lim x^{-m} = (\lim x)^{-m}. \qquad (2.4)$$

3. Сравнение бесконечно малых величин. Отношение двух бесконечно малых величин α и β может быть либо бесконечно большой, либо бесконечно малой, либо конечной величиной, либо величиной, не имеющей предела. В общем виде об отношении двух бесконечно малых величин нельзя сказать что-либо определенное. Это отношение зависит от характера изменения бесконечно малых величин.

Установить характер отношения двух бесконечно малых величин или, как говорят, раскрыть неопределенность вида 0/0 можно путем определения предела отношения α/β. Здесь имеют место следующие ситуации.

1. Предположим, что имеются бесконечно малые величины $\alpha = x^3$ и $\beta = x$, тогда

$$\lim \frac{\alpha}{\beta} = \frac{x^3}{x} = 0.$$

Величина $\alpha = x^3$, стремящаяся к нулю быстрее, чем $\beta = x$, выступает в качестве бесконечно малой *высшего порядка* малости относительно β.

2. Если, например, $\alpha = x$, а $\beta = x^2$, то

$$\lim \frac{\alpha}{\beta} = \frac{x}{x^2} = \infty.$$

В этом случае α является бесконечно малой *низшего порядка* малости относительно β.

3. Предел отношения может быть равен конечному числу, например, в случае, когда $\alpha = (x + 5)\,x$, а $\beta = x$:

$$\lim \frac{\alpha}{\beta} = \lim_{x \to 0} \frac{(x+5)\,x}{x} = \lim_{x \to 0} (x+5) = 5.$$

Здесь бесконечно малые α и β являются величинами *одного и того же порядка малости*.

4. Отношение α/β может вовсе не иметь предела, как, например, в случае, если $\alpha = x \sin \frac{1}{x}$, а $\beta = x$. Предел

$$\lim_{x \to 0} \frac{x \sin \frac{1}{x}}{x} = \lim_{x \to 0} \sin \frac{1}{x}$$

не существует.

5. Предел отношения бесконечно малых величин α и β может быть равен единице, т. е.

$$\lim \frac{\alpha}{\beta} = 1.$$

В этом случае α и β называются *эквивалентными* бесконечно малыми величинами.

§ 2. Число Эйлера и модель роста словаря

В различных областях математики и ее приложениях часто используется функция

$$z = \left(1 + \frac{1}{x}\right)^x,$$

пределом которой при неограниченном возрастании x является уже встречавшееся нам число Эйлера $e = 2{,}718\ldots$ (см. гл. 1, § 6, п. 3). Таким образом, можно записать, что

$$e = \lim_{x \to \infty} \left(1 + \frac{1}{x}\right)^x. \qquad (2.5)$$

С помощью предела (2.5) решаются многие задачи квантитативной лингвистики, связанные с ростом или убыванием какой-либо величины.

Рассмотрим в этой связи некоторую идеальную модель роста словаря.

В результате постоянного расширения сферы деятельности человека лексика каждого языка, особенно его терминологический словарь, несмотря на выпадение некоторого количества слов, неуклонно растет. Характеристикой увеличения словаря служит коэффициент k его прироста за определенный период времени, например, за десятилетие. Этот коэффициент представляет собой отношение количества новых слов, появившихся за десятилетие, за вычетом вышедших из употребления архаизмов (ΔL), к общему объему (L) словаря в данный период, т. е.

$$k = \frac{\Delta L}{L}.$$

Зная начальный объем словаря L и коэффициент k, легко показать, что через 10 лет объем словаря составит

$$L_0 + \Delta L = L_0 + L_0 k = L_0 (1 + k).$$

Однако этот подсчет является в значительной степени приближенным. Ведь в течение десятилетия прирост словаря происходит не относительно исходной величины L_0, а относительно сумм $L_0 + \Delta L$ (здесь величина ΔL указывает прирост словаря за то количество времени — за год, месяц, неделю, — которое прошло от момента, когда был зафиксирован начальный объем словаря L_0).

Предположим, что нам нужно учесть рост словаря по годам. Тогда к концу первого года мы получим объем словаря, равный

$$L_1 = L_0 \left(1 + \frac{k}{10}\right),$$

к концу второго года он будет равен

$$L_2 = L_1 + L_1 \frac{k}{10} = L_0 \left(1 + \frac{k}{10}\right)^2,$$

а к концу десятилетия объем словаря составит

$$L_{10} = L_0 \left(1 + \frac{k}{10}\right)^{10}.$$

Если же учитывать прирост словаря не по годам, а по месяцам, то получим еще более точный результат: к концу десятилетия объем словаря будет равен

$$L_0 \left(1 + \frac{k}{120}\right)^{120}.$$

Теперь попытаемся представить процесс роста словаря в общем виде. Для этого будем определять изменение словаря относительно промежутка времени T (например, тысячелетия), считая, что начальный объем словаря характеризуется величиной L_0, а коэффициент прироста по-прежнему равен k. Разделим весь период T на n малых равных частей:

$$\left[0, \frac{T}{n}\right], \left[\frac{T}{n}, \frac{2T}{n}\right], \ldots, \left[\frac{(i-1)T}{n}, \frac{iT}{n}\right], \ldots, \left[\frac{(n-1)T}{n}, T\right].$$

Поскольку все промежутки $\frac{iT}{n} - \frac{(i-1)T}{n} = \frac{T}{n}$ малы, то в каждом из них можно считать прирост новых слов ΔL постоянным и пропорциональным исходному объему словаря и величине промежутка T/n, т. е. $\Delta L = kL_0 T/n$.

Объем словаря к концу первого промежутка составит

$$L_1 = L_0 + kL_0 \frac{T}{n} = L_0 \left(1 + \frac{kT}{n}\right),$$

к концу второго промежутка

$$L_2 = L_0 \left(1 + \frac{kT}{n}\right)^2,$$

и, наконец, к концу эпохи T словарь будет включать

$$L_T = L_0 \left(1 + \frac{kT}{n}\right)^n$$

слов.

Предполагая, что число промежутков неограниченно растет (т. е. $n \to \infty$), а длина промежутка T/n неограниченно убывает (т. е. $T/n \to 0$), являясь тем самым бесконечно малой величиной, приходим к равенству

$$L_T = \lim_{n \to \infty} L_0 \left(1 + \frac{kT}{n}\right)^n. \qquad (2.6)$$

Заменим частное kT/n величиной $1/x$, тогда $n = kTx$. Поскольку kT/n представляет собой произведение постоянной величины k на бесконечно малую величину T/n, то $1/x$ есть тоже бесконечно малая величина, т. е. $1/x = kT/n \to 0$. Учитывая все сказанное, имеем:

$$L_T = \lim_{x \to \infty} L_0 \left(1 + \frac{1}{x}\right)^{kTx} = \lim_{x \to \infty} L_0 \left[\left(1 + \frac{1}{x}\right)^x\right]^{kT}.$$

Применяя следствия 1 и 2 из теоремы 2 (см. § 1, п. 2), перепишем это выражение в виде

$$L_T = L_0 \left[\lim_{x \to \infty} \left(1 + \frac{1}{x}\right)^x\right]^{kT},$$

откуда на основании соотношения (2.5) получаем формулу объема словаря к концу периода T в виде

$$L_T = L_0 e^{kT}. \qquad (2.7)$$

§ 3. Глоттохронология

1. Классическая глоттохронология. На только что рассмотренный математический аппарат опирается современная глоттохронология, целью которой является приближенная абсолютная датировка процессов расхождения диалектов и родственных языков, а также количественная оценка степени их родства.

Исходные лингвистические предпосылки глоттохронологии [32в, с. 384] сводятся к пяти постулатам.

1⁰. Во всех языках мира существует некоторое множество слов, обозначающих наиболее древние, всегда необходимые и поэтому не изменяющиеся понятия-означаемые (например: *большой*, *все*, *дерево*, *птица* и т. д.). Их можно объединить в некоторый тестовый список (ТС).

2⁰. Доля означающих (слов или эквивалентных словам устойчивых словосочетаний) из ТС, которая сохраняется за некоторый промежуток времени T для каждого языка, постоянна и не зависит от способа выбора этих слов из ТС.

3⁰. Каждый язык и диалект имеет свой коэффициент сохранности r относительно периода в 1000 лет. Величины r колеблются относительно периода в тысячу лет от 0,74 (для быстро развивающихся «динамичных» языков) до 0,91 (для «стабильных» языков), средняя величина $\bar{r}$ равна 0,81.

4⁰. Все означающие из ТС данного языка имеют одинаковые шансы сохраниться на протяжении интервала T.

5⁰. Если мы имеем дело с двумя потомками некоторого праязыка, то вероятность означающего из праязыкового ТС удержаться в ТС первого потомка не зависит от вероятности сохраниться в ТС второго потомка.

Опираясь на эти постулаты, можно, во-первых, математически оценить число тех означающих из ТС праязыка, которое сохраняется в двух языках-потомках за период их независимого развития; во-вторых, получить количественную оценку времени их самостоятельного развития.

Предположим, что L_0 есть известное нам число означающих ТС, которые характеризуют праязык в момент выделения из него языков-потомков i и j. Зная коэффициенты сохранности лексики для каждого из сравниваемых языков (r_i и r_j), легко определить коэффициент потери общих слов в ходе дивергенции. Этот коэффициент равен $k = 1 - r_i r_j$. Теперь, опираясь на рассуждения, приведенные в § 2, легко показать, что число общих означающих, сохранившихся в языках-потомках за T тысячелетий их самостоятельного развития, будет равно

$$L_T = L_0\left(1 - \frac{kT}{n}\right)^n = L_0\left[1 + \frac{(-k)\,T}{n}\right]^n \qquad (2.8)$$

(напомним, что n — произвольное количество отрезков, на которое разделено время дивергенции T). Повторяя преобразования, использованные в § 2, приходим к равенству

$$L_T = L_0 e^{-kT}. \qquad (2.9)$$

Зависимость (2.9) показывает, сколько слов из ТС, определенного в момент распада праязыка T_0, доживет до момента T_1 (здесь $T_1 - T_0 = T$).

Степень родства языков и диалектов можно оценивать через отношение $\gamma = L_T/L_0$. При $\gamma \geqslant 0{,}81$ два «говорения» следует считать диалектами, при $0{,}36 \leqslant \gamma < 0{,}81$ «говорения» являются родственными языками, входящими в одну семью; при $0{,}12 \leqslant \gamma < 0{,}36$ их следует считать принадлежащими к одной ветви.

Из соотношения (2.9) легко получить значение периода дивергенции T. Для этого, прологарифмировав (2.9), получаем

$$\ln L_T = \ln L_0 - kT,$$

откуда приходим к равенству

$$T = \frac{\ln L_0 - \ln L_T}{k}. \qquad (2.10)$$

Теперь, опираясь на данные, приводимые В. Гуцу-Ромало [50, с. 576—584], определим период дивергенции для пары романских «говорений» — дакорумынского и арумынского языков-диалектов. Всего сравнивается $L_0 = 202$ латинских слова, из которых оба языка-диалекта сохраняют только $L_T = 149$ общих лексем. Коэффициент сохранности лексики для дакорумынского языка-диалекта составляет $r_d = 0{,}81$, для арумынского варианта соответственно имеем $r_a = 0{,}88$ [32в, с. 386].

Тогда коэффициент потери общей лексики равен

$$k = 1 - r_d r_a = 1 - 0{,}81 \cdot 0{,}88 \approx 0{,}29.$$

С помощью этих данных на основании формулы (2.10) вычисляем период дивергенции для обоих языков-диалектов. Этот период составляет

$$= \frac{\ln 202 - \ln 149}{0{,}29} = 1{,}05 \text{ тысячелетия}.$$

Иными словами, распадение балканороманского языка-основы относится к началу IX в., что хорошо согласуется с историко-лингвистическими фактами [24, с. 273—278].

2. Ранговый метод в глоттохронологии. Классическая глоттохронология имеет ряд уязвимых пунктов.

Во-первых, критерии отбора понятий для ТС не являются простыми, объективными и однозначными. Сами значения используемые в ТС, неизоморфны для разных языков. Примером может служить неконгруэнтность означаемых у прилагательных, обозначающих цвета, у терминов родства, у существительных со значением «дерево» [32 в, с. 285]; [44]. Поэтому между ТС и словарем конкретного языка трудно установить однозначное соответствие.

Во-вторых, нет никакой уверенности в том, что степень сохранности лексики одинакова для всех участков ТС (ср. постулаты 2⁰ и 4⁰).

В-третьих, классическая глоттохронология не учитывает возможности вторичного сближения родственных языков, так широко представленного в индоевропейском, тюркском и других ареалах.

Преодолеть некоторые из этих трудностей стремятся М. В. Арапов и М. М. Херц [4], пытающиеся, во-первых, распространить глоттохронологическую методику на лексику, находящуюся за пределами ТС, а во-вторых, учесть различную скорость изменения разных групп слов.

Отправной точкой построения авторов является идея о том, что ранг (частота) слова в частотном словаре и его возраст коррелированы: чем чаще употребляется слово и чем меньше его ранг, тем больше вероятность того, что это слово древнего происхождения. В связи с этим предполагается, что группы слов с близкими рангами (частотами) в частотном словаре (ЧС) ведут себя так же, как и слова ТС. Такое предположение дает возможность авторам отказаться от использования узкого тестового списка и осуществлять исследование на словарном материале неограниченного объема.

Если в частотном словаре выбрать L_0 лексических единиц (слов, словоформ, основ, словосочетаний) с номерами $i + L_0$, то число слов, имеющих в промежутке $[i, i + L_0]$ возраст не менее T лет, вычисляется по формуле

$$L_T^* = L_0 e^{-k^* T}, \qquad (2.11)$$

где k^* — постоянная, указывающая, в каком месте ЧС находится отрезок $[i, i + L_0]$.

Чтобы учесть зависимость между скоростью изменения группы $[i, i + L_0]$ и местом ЧС, авторы разбивают ЧС на последовательные группы одинаковой длины по L_0 слов, присваивая каждой группе определенный номер. При этом выясняется, что существует зависимость

$$k^* = \alpha \sqrt{j}, \qquad (2.12)$$

где j — номер группы, α — параметр, характеризующийся объемом группы L_0 и тем, на каких лексических единицах построен ЧС.

Подставляя в (2.11) значение k^* из (2.12), авторы приходят к основной зависимости их теории:

$$L_{T,j} = L_0 e^{-\alpha T \sqrt{j}}. \qquad (2.13)$$

Зафиксируем возраст слов T и будем считать произведение αT параметром η формулы (2.13). Тогда доля слов, употребляющихся в языке не менее T тысячелетий, быстро убывает с ростом ранга группы:

$$f_{T,j} = e^{-\eta \sqrt{j}}. \qquad (2.14)$$

Проверка зависимости (2.14) на материале ЧС русского языка [39] показала хорошее схождение теоретических и опытных данных. Об этом свидетельствует рис. 27, на котором верхняя теоретическая прямая и эмпирические точки соответствуют синхронному срезу 1500 г., а нижняя — срезу 600 г. (праславянская эпоха). По оси абсцисс выбран масштаб корня квадратного, по оси ординат — логарифмический масштаб. График заимствован из работы [4].

Если фиксировать ранг j и вернуться к параметру k^*, то можно прийти к зависимости

$$f_{T,j} = e^{-k^* T},$$

связанной с исходными формулами глоттохронологии (2.9) и (2.11). Таким образом, соотношение (2.13) выступает в качестве обобщения зависимости (2.9).

Исходя из равенства (2.13), можно построить формулу, оценивающую число лексических единиц, имеющих возраст меньше T тысячелетий, т. е. появившихся после некоторого момента. Это число равно

$$L_{T,j} = L_0 \left(1 - e^{-\alpha T \sqrt{j}}\right) \qquad (2.15)$$

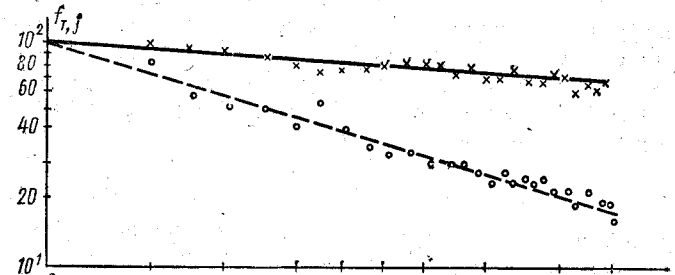


Рис. 27

и будет увеличиваться по мере роста ранга группы j, т. е. по мере того как мы будем обращаться ко все более редким словам.

Затем авторы вводят зависимость

$$L_{T,\,T+\Delta T,\,j} = L_{T,j} - L_{T+\Delta T,\,j} = L_0 \left(e^{-\alpha T \sqrt{j}} - e^{-\alpha (T+\Delta T) \sqrt{j}}\right), \qquad (2.16)$$

оценивающую число слов, появившихся в языке в период от момента T до момента $T + \Delta T$.

Исследование выражения (2.16) показывает, что число неологизмов, появившихся, например, за 500 лет ($\Delta T = 0{,}5$) при коэффициенте $\alpha = 0{,}02$, по мере перехода ко все более редким словам сначала растет, достигая максимума при $j = 1600$, что соответствует рангу $i = 160\,000$ (если принять объем порции равным ста словам), затем число неологизмов начинает падать.

Приведенный математический аппарат интересен с точки зрения перспектив математического моделирования различных диахронических процессов. Действительно, если, например, выделить в языке A слова, затронутые действием какого-либо фонетического, морфологического или словообразовательного закона, то всегда можно найти экспериментальное распределение каждого из этих классов по ранговым группам ЧС.

Если бы удалось аппроксимировать эти распределения с помощью выражений (2.13), (2.15), (2.16), то лингвисты получили бы

возможность количественно оценивать степень исходного и вторичного лексического родства двух языков. Одновременно появилась бы возможность получать абсолютную хронологию указанных диахронических процессов.

§ 4. Информационные модели слова и текста

Рассмотренный в § 2 и 3 математический аппарат может быть применен для моделирования не только диахронических, но и информационных процессов в речи. В частности, с помощью вышеприведенных рассуждений можно получить строгое доказательство формулы контекстной обусловленности, а также формул распределения информации в тексте и слове.

Опыты по угадыванию букв неизвестного текста или слова показывают, что энтропия (неопределенность) H перед угадыванием каждой последующей буквы (а, следовательно, и получаемая в результате угадывания информация I) последовательно убывает. Получены экспериментальные оценки коэффициента убывания энтропии для некоторых разновидностей русского и французского языков. Предположим теперь, что, зная начальную неопределенность H_0 и коэффициент убывания s, мы должны определить теоретическую неопределенность буквы, стоящей в самом конце слова длиной в ξ букв (ξ — величина непрерывная). Эту неопределенность мы обозначим символом H_ξ. Повторяя приведенные выше рассуждения, разделим весь текст на m равных участков

$$\left[0, \frac{\xi}{m}\right], \left[\frac{\xi}{m}, \frac{2\xi}{m}\right], \ldots, \left[\frac{(i-1)\xi}{m}, \frac{i\xi}{m}\right], \ldots, \left[\frac{(m-1)\xi}{m}, \xi\right].$$

Если полученные участки $\frac{i\xi}{m} - \frac{(i-1)\xi}{m} = \frac{\xi}{m}$ достаточно малы, то в каждом таком промежутке абсолютную величину убывания энтропии $|\Delta H|$ можно считать постоянной и в то же время пропорциональной начальной энтропии H_0 и ширине участка ξ/m, т. е. $|\Delta H| = sH_0\xi/m$.

Количество энтропии в момент ξ/m равно

$$H_1 = H_0 - sH_0\frac{\xi}{m} = H_0\left(1 - \frac{s\xi}{m}\right),$$

в момент $2\xi/m$ оно составляет

$$H_2 = H_1 - sH_1\frac{\xi}{m} = H_1\left(1 - \frac{s\xi}{m}\right) = H_0\left(1 - \frac{s\xi}{m}\right)^2,$$

и, наконец, в конце текста ξ энтропия будет равна

$$H_\xi = H_0\left(1 - \frac{s\xi}{m}\right)^m.$$

Снова предположив, что число участков неограниченно растет ($m \to \infty$), а их длина неограниченно убывает ($\xi/m \to 0$), получим

$$H_\xi = \lim_{m\to\infty} H_0\left(1 - \frac{s\xi}{m}\right)^m.$$

Заменив величину $-s\xi/m$ дробью $1/x$ (при этом $m = -xs\xi$) и учитывая следствия 1 и 2 из теоремы 2 (см. § 1, п. 2), находим

$$H_\xi = \lim_{x\to\infty} H_0\left(1 + \frac{1}{x}\right)^{-xs\xi} = H_0\left[\lim_{x\to\infty}\left(1 + \frac{1}{x}\right)^x\right]^{-s\xi}.$$

Так как предел, стоящий в квадратных скобках, есть число Эйлера, то

$$H_\xi = H_0 e^{-s\xi}. \tag{2.17}$$

Заменив непрерывную величину, характеризующую длину слова, на дискретные целочисленные значения буквенных позиций n и учитывая, что неопределенность H количественно равна синтактической информации I (см. гл. 5, § 5, п. 2), приходим к выражению

$$I_n = I_0 e^{-sn}, \tag{2.18}$$

характеризующему количество информации, которое несет буква, стоящая на n-м месте в слове.

Теперь используем формулу (2.17) для анализа распределения информации в тексте. Нетрудно заметить, что при бесконечном увеличении длины текста ($\xi \to \infty$) величина $e^{-s\xi}$ стремится к нулю, поэтому $\lim_{\xi\to\infty} H_0 e^{-s\xi} = 0$, а следовательно, и $\lim_{\xi\to\infty} H_\xi = 0$.

Однако информационные исследования текста [23]; [26] показывают, что бесконечное увеличение его длины не приводит к полной утрате неопределенности продолжений. Действительно, всякий текст, будучи образован из сложных знаков (слов, словосочетаний, предложений), обладающих практически неограниченной комбинаторной способностью, имеет несколько продолжений или, иначе говоря, всегда обладает неопределенностью выбора следующей лингвистической единицы. Таким образом, предельная энтропия H_∞ всегда больше нуля. Из всего сказанного следует, что величина H_∞, определяемая обычно из опыта, должна быть исключена из нашего расчета. Таким образом, моделируя распределение информации в тексте с помощью выражения (2.17), мы должны оперировать вместо H_ξ разностью $H_\xi - H_\infty$, а вместо H_0 — разностью $H_0 - H_\infty$. Учитывая это, имеем

$$H_\xi - H_\infty = (H_0 - H_\infty)\, e^{-s\xi},$$

откуда получаем

$$H_\xi = (H_0 - H_\infty)\, e^{-s\xi} + H_\infty.$$

Заменив величины H на значения I и введя вместо непрерывного ξ дискретное n, приходим к формуле распределения информации в связном тексте:

$$I_n = (I_0 - I_\infty)\, e^{-sn} + I_\infty. \tag{2.19}$$

Если в гл. 1 эта формула имела вид более или менее удачно подобранной аппроксимации опытных наблюдений, то теперь, опираясь на аппарат теории пределов и бесконечно малых величин, мы получили строгое ее доказательство, раскрывающее внутренний процесс построения информационной схемы слова и предложения.

ГЛАВА 3

ДИНАМИКА ЛИНГВИСТИЧЕСКИХ ПРОЦЕССОВ И ЕЕ ОПИСАНИЕ С ПОМОЩЬЮ ПРИЕМОВ ДИФФЕРЕНЦИАЛЬНОГО ИСЧИСЛЕНИЯ

§ 1. Диахроническая скорость и понятие производной

В предыдущих главах мы познакомились с простейшими математическими моделями, которые используются при описании диахронии языка и информационных процессов создания текста. Для того чтобы перейти к построению более сложных моделей, необходимо уметь измерять динамику лингвистического процесса на различных его этапах. При измерении этой динамики используется аппарат дифференциального исчисления, одним из исходных понятий которого является понятие производной. Чтобы раскрыть это понятие, приведем два лингвистических примера.

1. Арабские заимствования в персидской прозе X—XII в. Французский востоковед Г. Лазар [32 в, с. 374], исследуя арабские заимствования в персидской прозе X—XII в., подсчитал число арабских слов и словоупотреблений в некоторых исторических и религиозных текстах этого периода (см. табл. 3.1).

Таблица 3.1

Доля арабизмов в персидских текстах X—XII в.

	Индексы текстов и время их написания							
	1(TB)	2(TT)	3(TC)	4(ZA)	5(TBq)	6(At)	7(TAf)	8(RS)
	963—964 г.	конец X в.	X—XI в.	1050 г.	1060 г.	1072—1073 г.	1145 г.	1200 г.
Длина текста (N)	2376	2302	2165	2289	2424	2299	2253	2259
Число арабизмов в тексте (N_a)	229	260	237	343	390	296	456	590
Объем словаря (L)	450	477	494	514	586	519	563	694
Число арабизмов в словаре (L_a)	109	152	128	195	224	183	288	357
Процент арабизмов в тексте $\left(\frac{N_a}{N}\cdot 100\%\right)$	9,64	11,29	10,94	14,98	16,09	12,87	20,24	26,11
Процент арабизмов в словаре $\left(\frac{L_a}{L}\cdot 100\%\right)$	24,22	31,86	25,91	37,94	38,22	35,26	51,15	51,44

Используя данные, приведенные в табл. 3.1, постараемся оценить темп проникновения арабизмов в персидскую прозу X—XII в.

Установим с помощью графика (рис. 28) зависимость между временем написания произведений и долей арабизмов, встречающихся в тексте и словаре. По оси абсцисс отложим даты создания произведений, а по оси ординат — доли арабизмов в тексте ($n = \frac{N_a}{N}\cdot 100\%$) и в словаре ($n' = \frac{L_a}{L}\cdot 100\%$) каждого из них.

Как следует из графика, увеличение употребления арабизмов в тексте и их нарастание в словаре аппроксимируется прямыми линиями, которым соответствуют линейные зависимости

$$y_L = a_L x + b_L \qquad (3.1)$$

(для арабизмов в словаре) и

$$y_N = a_N x + b_N \qquad (3.2)$$

(для арабизмов в тексте); здесь y — доля арабизмов; x — время, начало отсчета которого соответствует началу X в. (т. е. 900 г.); a и b — параметры зависимости.

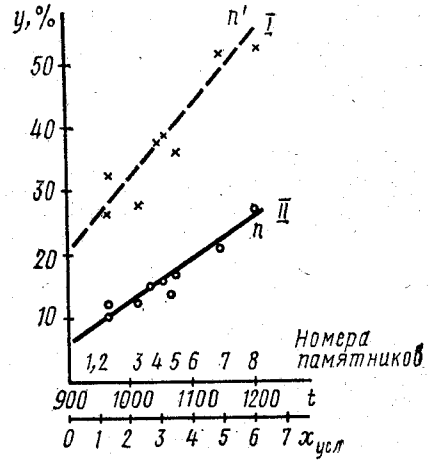


Рис. 28

Как известно, линейная зависимость рассматриваемого типа описывает равномерное движение некоторой материальной точки. Воспользовавшись физической аналогией, условимся считать такой материальной точкой долю арабизмов, а пройденный точкой путь y — количественным ростом этой доли в процентах.

Предположим теперь, что к моменту времени $x = 0$ наша точка уже прошла путь y_0, а в последующий момент времени x длина пути, пройденного точкой, равна y. Тогда расстояние, пройденное точкой за время x, равно разности $y - y_0$, а скорость, согласно определению равномерного движения, выразится отношением

$$v = (y - y_0)/x. \qquad (3.3)$$

Из равенства (3.3) нетрудно получить следующее выражение:

$$y = vx + y_0. \qquad (3.4)$$

Соотношение (3.4) описывает закон равномерного движения и представляет собой линейную зависимость относительно x. При этом v (скорость равномерного движения) есть величина постоянная. В нашем примере v характеризует темп диахронического процесса, точнее, скорость увеличения доли арабизмов в персидском словаре или тексте. В дальнейшем величину v будем называть *диахронической скоростью*.

Возьмем теперь некоторый момент времени x_1. «Путь», пройденный долей арабизмов к моменту времени x_1, определим из выражения (3.4); он равен

$$y_1 = vx_1 + y_0.$$

Аналогично получим «путь», пройденный арабизмами к моменту времени x_2:

$$y_2 = vx_2 + y_0.$$

Следовательно, за отрезок времени $x_2 - x_1$ «точка» пройдет «путь»

$$y_2 - y_1 = (vx_2 + y_0) - (vx_1 + y_0),$$

или

$$y_2 - y_1 = v(x_2 - x_1),$$

откуда получаем

$$(y_2 - y_1)/(x_2 - x_1) = v. \qquad (3.5)$$

Как было отмечено, «путь», пройденный лингвистической «точкой» при равномерном диахроническом движении, есть функция от времени, т. е. $y = f(x)$. В нашем примере аргумент последовательно принимает два значения: x_1 и x_2. Разность $x_2 - x_1$ называется *приращением аргумента* и обозначается символом Δx (читается «дельта x»):

$$\Delta x = x_2 - x_1. \qquad (3.6)$$

Разность значений функций, соответствующих значениям аргумента x_1 и x_2, называется *приращением функции* и обозначается символом Δy:

$$\Delta y = f(x_2) - f(x_1) = y_2 - y_1. \qquad (3.7)$$

На основании введенных понятий приращения аргумента и функции выражение (3.5) может быть записано в виде

$$\frac{\Delta y}{\Delta x} = v. \qquad (3.8)$$

Иными словами, скорость равномерного диахронического движения есть отношение приращения «пути» к соответствующему приращению времени.

Определим теперь числовые значения приращений аргумента (времени) и функции (доли арабизмов), а также скорости проникновения арабизмов в персидскую прозу.

На основании графика (см. рис. 28) составим таблицу теоретических значений долей арабизмов в словаре персидского литературного языка разных периодов. Условной единицей времени $x_{усл}$ будем считать промежуток в 50 лет, а величина y выражается в процентах.

Используя данные табл. 3.2, определим по формуле (3.8) среднюю скорость v в различные моменты времени x. Результаты приведены в табл. 3.3.

Нетрудно видеть, что диахроническая скорость проникновения арабизмов в персидскую прозу X—XII в. является постоянной величиной и составляет 6% за единицу времени (т. е. за 50 лет).

Таблица 3.2

T (годы)	900	950	1000	1050	1100	1150	1200
$x_{усл}$	0	1	2	3	4	5	6
y (%)	20	26	32	38	44	50	56

Таблица 3.3

i	x_i	$y(x_i)$	$x_{i+1} - x_i = \Delta x$	$y(x_{i+1}) - y(x_i) = \Delta y$	$\frac{\Delta y}{\Delta x} = v$
1	0	20	1	6	6
2	1	26	1	6	6
3	3	32	1	6	6
4	3	38	1	6	6
5	4	44	1	6	6

2. История употребления местоимения hic в позднелатинских памятниках. Мгновенная диахроническая скорость и понятие производной. Может показаться, что мы слишком усложнили дело, прибегая к сравнению приращения лингвистического «пути» с приращением времени. Гораздо проще было бы, определяя диахроническую скорость, поделить «путь», пройденный долей арабизмов, на время x. Однако рассмотренная задача решается подобным образом только в том случае, когда мы имеем дело с равномерно развивающимся лингвистическим процессом, который можно смоделировать при помощи линейной функции. Между тем историческому языкознанию и лингвистике речи приходится иметь дело с такими процессами, в которых лингвистическая скорость не является постоянной величиной. Рассмотрим, в частности, историю употребления латинского местоимения hic.

Судя по данным классической прозы и позднелатинским памятникам, латинское hic было наиболее вероятным претендентом среди других латинских местоимений на роль формировавшегося в то время романского артикля. Латинские памятники первых веков нашей эры показывают постепенное возрастание употребительности форм hic. Однако примерно в IV в. этот процесс прекращается и употребительность hic идет на убыль. Это отступление отмечается и в латинских документах раннего средневековья, отражающих протороманскую речь. Имеющиеся статистические данные об употреблении hic [27, с. 12—15]; [32 в, с. 377—381] хорошо аппроксимируются следующей формулой:

$$y = 23 - 0{,}75x^2 + 6x, \qquad (3.9)$$

где y — доля употребительности hic в % среди других указательных местоимений, а x — время, измеряемое в условных единицах (каждая единица составляет столетие).

Используя формулу (3.9), определим теоретическую долю употребительности hic в самом начале нашей эры («нулевой» год), в 100-м, 200-м и т. д. годах н. э., а также увеличение этой доли за каждые 100 лет. Все эти данные показаны в табл. 3.4. На их основе построен график, изображенный на рис. 29.

Таблица 3.4

i	x_i	y (%)	$y_{i+1}-y_i=v_{ср}$ (ед. вр.)
0	(начало н. э.)	23,00	5,25
1	(100 г.)	28,25	3,75
2	(200 г.)	32,00	2,25
3	(300 г.)	34,25	0,75
4	(400 г.)	35,00	—0,75
5	(500 г.)	34,25	—2,25
6	(600 г.)	32,00	—3,75

Из табл. 3.4 видно, что за каждую условную единицу времени (столетие) происходит увеличение доли употребительности hic ($y_{i+1} - y_i$), причем этот прирост осуществляется неравномерно:

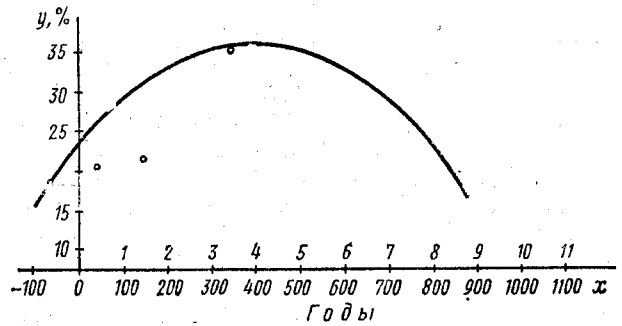


Рис. 29

в первую единицу времени (до 100 г. н. э) эта доля увеличилась на 5,25%, между 100-м и 200-м годами уже только на 3,75% и т. д, а после 400 г. вообще отмечается падение употребительности hic.

Прирост доли hic за 100 лет можно рассматривать как среднюю скорость ($v_{ср}$) нарастания роста его употребительности в течение нашей условной единицы времени (т. е. в столетие). Эта скорость не является постоянной величиной. Но если скорость роста употребительности hic, равная 5,25% в столетие в течение первого промежутка, упала до 3,75% в столетие в течение второго промежутка, то, следовательно, она менялась и внутри этих промежутков.

Очевидно, что если брать промежутки времени все меньшей и меньшей величины, то можно в конце концов прийти к мгновенной скорости. Определим мгновенную скорость нарастания употреби-

Таблица 3.5

x	2 (200 г.)	2,001 (6.II 200 г.)	2,01 (201 г.)	2,1 (210 г.)	2,3 (230 г.)	2,5 (250 г.)
y (%)	32	32,0030	32,0299	32,2925	32,8625	33,3125

тельности hic в самом начале 200-го г. н. э. Для этого, получив значения y для разных моментов времени в интервале от $x = 2$ до $x = 2{,}5$ (см. табл. 3.5), с помощью выражения (3.9) будем вычислять среднюю скорость для все более малых промежутков времени. В итоге получим следующие результаты:

от 2 до 2,5 имеем $v_{ср} = \frac{33{,}3125-32}{0{,}5} = 2{,}625\%$ в столетие;

» 2 » 2,3 » $v_{ср} = \frac{32{,}8625-32}{0{,}3} = 2{,}875\%$ » » ;

» 2 » 2,1 » $v_{ср} = \frac{32{,}2925-32}{0{,}1} = 2{,}925\%$ » » ;

» 2 » 2,01 » $v_{ср} = \frac{32{,}0299-32}{0{,}01} = 2{,}990\%$ » » ;

» 2 » 2,001 » $v_{ср} = \frac{32{,}0030-32}{0{,}001} = 3{,}000\%$ » » .

Из полученных данных видно, что чем меньше промежуток времени, тем ближе значение средней скорости к 3%. Короче говоря, можно утверждать, что мгновенная диахроническая скорость в момент $x = 2$ составляет 3% в столетие.

Теперь определим мгновенную скорость в момент $x = 3$ (см. табл. 3.6). Характер вычислений тот же, что и в предыдущем примере.

Таблица 3.6

x	3	3,001	3,01	3,1	3,2	3,3
y (%)	34,25	34,2515	34,2649	34,3925	34,5200	34,6325

Таким образом,

от 3 до 3,3 имеем $v_{ср} = \frac{34{,}6325-34{,}25}{0{,}3} = 1{,}275\%$ в столетие;

» 3 » 3,2 » $v_{ср} = \frac{34{,}5200-34{,}25}{0{,}2} = 1{,}350\%$ » » ;

от 3 до 3,1 имеем $v_{ср} = \frac{\approx 4,3925 - 34,25}{0,1} = 1,400\%$ в столетие;

» 3 » 3,01 » $v_{ср} = \frac{34,2649 - 34,25}{0,01} = 1,493\%$ » » ;

» 3 » 3,001 » $v_{ср} = \frac{34,2515 - 34,25}{0,001} = 1,5\%$ » » .

Итак, можно считать, что мгновенная скорость в момент $x = 3$ составляет 1,5% в столетие.

Рассматривая равномерный лингвистический процесс, мы пришли к выводу, что его скорость есть величина постоянная и она остается неизменной при любом значении x и при любом значении Δx. Иначе обстоит дело при неравномерном процессе. На примере развития употребительности латинского местоимения hic из произведенных нами расчетов видно, что $v_{ср}$ зависит и от момента времени x, и от Δx. Очевидно, что при одном и том же значении Δx различным моментам времени x соответствуют различные значения $\frac{\Delta y}{\Delta x}$. С другой стороны, при одном и том же значении x величина отношения $\frac{\Delta y}{\Delta x}$ зависит от Δx, причем чем меньше промежуток Δx, тем ближе величина скорости к мгновенной.

Поэтому при неравномерном движении можно говорить не о скорости материальной точки вообще, а только о скорости в данный момент или, как говорилось выше, о мгновенной скорости. Отсюда следует, что величину $v(x)$ мгновенной скорости можно рассматривать как предел отношения $\frac{\Delta y}{\Delta x}$ при Δx, стремящемся к нулю, т. е.

$$v(x) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}. \qquad (3.10)$$

Вполне очевидно, что функция $v(x)$ зависит от функции $y(x)$: функция $y(x)$ как бы производит функцию $v(x)$. Поэтому говорят, что функция $v(x)$ является *производной* функции $y(x)$.

3. Нахождение производной. Из всего сказанного в п. 1 и 2 следует, что *производной функции* $y = f(x)$ является *предел отношения приращения* Δy *этой функции к приращению независимой переменной* Δx *при стремлении* Δx *к нулю*, т. е.

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}. \qquad (3.11)$$

Для обозначения производной функции обычно используются символы y' или y'_x или $f'(x)$ (читается: «игрек штрих», «игрек штрих по икс», «эф штрих по икс»). Нахождение производной y'_x функции $y = f(x)$ по аргументу x, или, иначе говоря, *дифференцирование*, можно разбить на следующие этапы:

1) нахождение приращения аргумента Δx и нового (наращенного) значения функции:

$$y + \Delta y = f(x + \Delta x);$$

2) определение приращения функции Δy:

$$\Delta y = f(x + \Delta x) - f(x);$$

3) нахождение отношения приращения функции к приращению аргумента:

$$\frac{\Delta y}{\Delta x} = \frac{f(x+\Delta x) - f(x)}{\Delta x};$$

4) определение предела этого отношения при условии $\Delta x \to 0$:

$$y'_x = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}.$$

Обращаем внимание читателя на то, что в рассмотренных выше примерах вычисление производной (мгновенной скорости диахронических процессов) осуществлялось по только что описанной схеме, которая носит название непосредственного дифференцирования. Однако непосредственное дифференцирование связано с громоздкими вычислениями. Поэтому при определении производной обычно применяют уже готовые стандартные правила и формулы, использование которых значительно упрощает процесс дифференцирования.

Ниже дается перечень таких правил и формул дифференцирования. Использующиеся в них символы расшифровываются следующим образом: x — аргумент; y — простая или сложная функция; u, v, w, z — сложные функции от x; a, c, n — постоянные величины; e — основание натуральных логарифмов.

Общие правила дифференцирования

$y = c$	$y'_x = 0$
$y = x$	$y'_x = 1$
$y = u \pm v \pm \ldots \pm w$	$y'_x = u'_x \pm v'_x \pm \ldots \pm w'_x$
$y = u \cdot v$	$y'_x = u'_x \cdot v + v'_x \cdot u$
$y = \frac{u}{v}$	$y'_x = \frac{u'_x \cdot v - v'_x \cdot u}{v^2}$
$y = \frac{u}{c}$	$y'_x = \frac{u'_x}{c}$
$y = \frac{c}{v}$	$y'_x = -\frac{cv'}{v^2}$
$y = f(u)$, где $u = \varphi(x)$	$y'_x = y'_u u'_x$

Производные основных элементарных функций

Степенная функция

$y = x^n$	$y'_x = nx^{n-1}$
$y = \sqrt{x}$	$y'_x = \frac{1}{2\sqrt{x}}$
$y = u^n$	$y'_x = nu^{n-1}u'_x$
$y = \sqrt{u}$	$y'_x = \frac{u'_x}{2\sqrt{u}}$

Показательная функция

$y = a^x$	$y'_x = a^x \ln a$
$y = a^u$	$y'_x = a^u \ln a \cdot u'_x$
$y = e^x$	$y'_x = e^x$
$y = e^u$	$y'_x = e^u u'_x$

Логарифмическая функция

$y = \log_a x$	$y'_x = \frac{1}{x}\log_a e$
$y = \ln x$	$y'_x = \frac{1}{x}$
$y = \log_a u$	$y'_x = \frac{u'_x}{u \ln a}$
$y = \ln u$	$y'_x = \frac{u'_x}{u}$

Тригонометрические функции

$y = \sin x$	$y'_x = \cos x$
$y = \sin u$	$y'_x = \cos u \cdot u'_x$
$y = \cos x$	$y'_x = -\sin x$
$y = \cos u$	$y'_x = -\sin u \cdot u'_x$
$y = \operatorname{tg} x$	$y'_x = \frac{1}{\cos^2 x}$
$y = \operatorname{tg} u$	$y'_x = \frac{u'_x}{\cos^2 u}$
$y = \operatorname{ctg} x$	$y'_x = -\frac{1}{\sin^2 x}$
$y = \operatorname{ctg} u$	$y'_x = -\frac{u'_x}{\sin^2 u}$

Обратные тригонометрические функции

$y = \arcsin x$	$y'_x = \frac{1}{\sqrt{1-x^2}}$
$y = \arcsin u$	$y'_x = \frac{u'_x}{\sqrt{1-u^2}}$
$y = \arccos x$	$y'_x = -\frac{1}{\sqrt{1-x^2}}$
$y = \arccos u$	$y'_x = -\frac{u'_x}{\sqrt{1-u^2}}$
$y = \operatorname{arctg} x$	$y'_x = \frac{1}{1+x^2}$
$y = \operatorname{arctg} u$	$y'_x = \frac{u'_x}{1+u^2}$
$y = \operatorname{arcctg} x$	$y'_x = -\frac{1}{1+x^2}$
$y = \operatorname{arcctg} u$	$y'_x = -\frac{u'_x}{1+u^2}$

Строгое доказательство правил и формул дифференцирования читатель найдет в книге [28].

§ 2. Дифференциал

1. Понятие дифференциала и дифференциалы простейших функций. В предыдущем параграфе было показано, что использование производной связано с нахождением приращения функции Δy. Так как последнее представляет собой довольно трудную задачу, то на практике пользуются более простым приближенным значением функции — ее дифференциалом. Эта идея лежит также в основе применения дифференциала к приближенным вычислениям.

Для разъяснения понятия дифференциала воспользуемся сначала конкретной задачей. Пусть соотношение между количеством терминов и терминологических словосочетаний, появляющихся в текстах интересующей нас науки, написанных на германских, романских и славянских языках, с одной стороны, и временем развития этой науки, с другой, описывается функцией

$$y = a + x^3, \tag{3.12}$$

где y — суммарное количество терминов в указанных языках, появившихся за x лет развития данной области знания, a—число слов и словосочетаний, использованных для обозначения исходных по-

нятий данной научной области. График этой функции изображен на рис. 30.

Тогда производная функции (3.12) согласно правилам дифференцирования равна

$$y'_x = 3x^2.$$

Предположим теперь, что интересующая нас наука существует уже сто лет и мы хотим узнать, на сколько увеличится словарь этой науки за два года. Это увеличение следует рассматривать как приращение функции

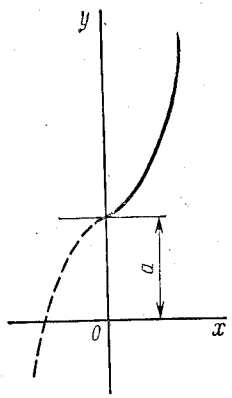


Рис. 30

$$\Delta y = (x + \Delta x)^3 - x^3.$$

Раскрывая скобки, получаем

$$\Delta y = x^3 + 3x^2\Delta x + 3x\Delta x^2 + \Delta x^3 - x^3,$$

или

$$\Delta y = 3x^2\Delta x + 3x\Delta x^2 + \Delta x^3.$$

Три члена, составляющие правую часть этого соотношения, не равноценны, в чем можно убедиться, подставляя численные значения $x = 100$ и $\Delta x = 2$:

$$3x^2\Delta x = 3 \cdot 10000 \cdot 2 = 60\,000;$$

$$3x\Delta x^2 = 3 \cdot 100 \cdot 4 = 1\,200; \quad \Delta x^3 = 2^3 = 8;$$

$$\Delta y = 61\,208.$$

Второе слагаемое меньше первого в 50 раз, а третье — в 7500 раз.

Таким образом, среди слагаемых, в сумме составляющих приращение Δy, величина $3x^2\Delta x$ является г л а в н о й частью приращения функции $f(x) = a + x^3$, в то время как сумма $3x\Delta x^2 + \Delta x^3 = \alpha$ составляет малозаметный вклад в величину приращения Δy. Поскольку слагаемое $3x^2\Delta x$ прямо пропорционально (при постоянном x) приращению Δx, его называют *главной линейной частью приращения* функции $f(x)$. Эта главная часть приращения, отличающаяся от последнего на бесконечно малую величину, и есть *дифференциал* функции:

$$dy = y'_x\Delta x. \qquad (3.13)$$

Так как приращение Δx аргумента совпадает с его дифференциалом dx, то отсюда следует, что дифференциал функции равен ее производной, умноженной на дифференциал аргумента, т. е.

$$dy = y'_x dx. \qquad (3.14)$$

Можно показать, что дифференциал функции и ее приращение суть эквивалентные бесконечно малые величины:

$$dy \approx \Delta y. \qquad (3.15)$$

Опираясь на равенство (3.14), мы приходим к новому важному для дальнейшего определению производной. Согласно этому определению, *производная функции есть отношение дифференциала этой функции к дифференциалу ее аргумента*:

$$y'_x = \frac{dy}{dx}.$$

Итак, для нахождения дифференциала функции следует найти производную этой функции и умножить ее на дифференциал аргумента, равный приращению последнего.

Применяя эту операцию, можно вывести формулы для нахождения дифференциалов простейших функций. Сводка этих формул приводится ниже (значения символов здесь те же, что и в сводке правил нахождения производной; см. § 1, п. 3).

$$da = 0; \qquad d(\arcsin u) = \frac{du}{\sqrt{1-u^2}};$$

$$dx = \Delta x;$$

$$d(u \pm v) = du \pm dv; \qquad d(\arccos u) = -\frac{du}{\sqrt{1-u^2}};$$

$$d(u \cdot v) = v \cdot du + u \cdot dv;$$

$$d\left(\frac{u}{v}\right) = \frac{v \cdot du - u \cdot dv}{v^2}; \qquad d(\operatorname{arc\,tg} u) = \frac{du}{1+u^2};$$

$$dy = y'_x \cdot dx; \qquad d(\operatorname{arc\,ctg} u) = -\frac{du}{1+u^2};$$

$$d(x^n) = n \cdot x^{n-1}\,dx; \qquad d(\log_a u) = \frac{du}{u \ln a};$$

$$d(\sin u) = \cos u\,du;$$

$$d(\cos u) = -\sin u\,du; \qquad d(a^u) = a^u \ln a\,du;$$

$$d(\operatorname{tg} u) = \frac{du}{\cos^2 u}; \qquad d(a^x) = a^x \ln a\,du;$$

$$d(\operatorname{ctg} u) = -\frac{du}{\sin^2 u}; \qquad d(e^u) = e^u\,du.$$

2. Использование дифференциала для приближенных вычислений в лингвистических задачах. Исследуя динамику диахронических процессов (ср. историю латинского местоимения hic в п. 2 § 1 и рост романо-германской и славянской терминологии) с помощью приращения функции Δy, мы видели, насколько трудоемки и громоздки вычисления этой величины, а следовательно, и самой функции. Вычисление значений функции y можно значительно упростить, опираясь на приближенное равенство

$$\Delta y \approx dy = y'_x\Delta x. \qquad (3.16)$$

Для этого перепишем выражение (3.16) в виде

$$f(x+\Delta x)-f(x)\approx y'_x\Delta x,$$

откуда

$$f(x+\Delta x)\approx f(x)+y'_x\Delta x,$$

или, учитывая (3.13),

$$f(x+\Delta x)\approx f(x)+df(x). \qquad (3.17)$$

Проверим теперь на оенове уже приводившихся данных о росте терминологии (см. п. 1) эффективность приближенных вычислений с помощью дифференциала.

Определим число терминов на 102-м году существования интересующей нас науки при условии, что на 100-м году ее развития имелось уже 1001000 терминов. Из них количество исходных терминов $a = 1000$, а вновь возникшие термины составляют $x^3 = 1000000$.

Примем за начальное значение аргумента $x = 100$, тогда $y = a + x^3 = 1001000$, $y'_x = 3x^2 = 30000$, а $\Delta x = 2$. Подставляя эти значения в (3.17), получаем

$$f(x+\Delta x) = f(102)\approx 1001000 + 30000\cdot 2 = 1061000.$$

Посмотрим теперь, какую ошибку мы допускаем, взяв дифференциал dy вместо приращения Δy. Для этого воспользуемся точным значением приращения нашей функции за два года:

$$\Delta y = 3x^2\Delta x + 3x\Delta x^2 + \Delta x^3 = 61208.$$

Точное число терминов через 102 года составит

$$y + \Delta y = 1001000 + 61208 = 1062208.$$

Допущенная ошибка при использовании dy вместо Δy равна

$$\frac{61208-60000}{1062208}\approx 0{,}00113,$$

т. е. всего лишь около 0,11%.

Используя равенство (3.17), можно получить приближенные формулы для вычисления значений некоторых элементарных функций. Приведем формулы, полезные в лингво-статистических и квантитативно-лингвистических исследованиях.

1. Если $y = x^n$, то $(x+\Delta x)^n \approx x^n + nx^{n-1}\Delta x$.

При $x = 1$ и $\Delta x = \alpha$ получим

$$(1+\alpha)^n \approx 1 + n\alpha; \qquad (3.18)$$

в случае $\Delta x = -\alpha$ выражение (3.18) принимает вид

$$(1-\alpha)^n \approx 1 - n\alpha. \qquad (3.19)$$

Мы будем применять равенство (3.19) при решении вероятностных и лингво-статистических задач. Пусть, например, необходимо возвести в десятую степень вероятность $q = 1 - P(A) = 0{,}9901$ (ср. гл. 6, § 1); тогда имеем:

$$q^{10} = (1-0{,}0099)^{10} = 1 - 10\cdot 0{,}0099 = 1 - 0{,}099 = 0{,}901.$$

Заметим, что формулой (3.19) следует пользоваться при небольших значениях n и α. Если $n > 10$, а $\alpha > 0{,}01$, то получаемые с помощью (3.19) значения функции будут заметно отличаться от истинных.

2. Если $y = e^x$, то, согласно (3.17),

$$e^{x+\Delta x} = e^x + (e^x)'_x\,\Delta x = e^x + e^x\,\Delta x. \qquad (3.20)$$

При определении количества информации, содержащейся в различных участках текста, возникает необходимость определить величину

$$f(n) = e^{-sn},$$

где $s = 0{,}21$ (см. табл. 1.5 на стр. 41), а n — номер буквенной позиции — аргумент функции $f(n)$.

Тогда можно вычислить теоретическое значение информации в каждой буквенной позиции беллетристического текста. Сделаем это относительно первой буквенной позиции.

Пользуясь формулой (3.17), имеем

$$f(n+\Delta n) = f(n) + f'(n)\,\Delta n,$$

т. е.

$$e^{-s(n+\Delta n)} = e^{-sn} - se^{-sn}\,\Delta n.$$

Далее, полагая $s = 0{,}21$, $n = 0$, а $\Delta n = 1$, получим

$$e^{-0{,}21} = e^0 - 0{,}21e^0\cdot 1 = 1 - 0{,}21 = 0{,}79.$$

Этот результат не очень сильно отличается от точного табличного значения $e^{-0{,}21} = 0{,}81$.

Аналогичным образом, пользуясь только что описанным методом, для второй буквенной позиции получаем $e^{-2\cdot 0{,}21} = 0{,}64$ при табличном значении 0,65.

3. Производные и дифференциалы высших порядков. Производную и дифференциал от данной функции $y = f(x)$ можно рассматривать как новые функции от x. Пользуясь правилами, приведенными выше, от этих функций можно взять новую производную и новый дифференциал.

Производная, взятая от производной $y'_x = f(x)$, называется *производной второго порядка* (или *второй производной*) и обозначается $(y'_x)'_x = y''_{xx}$ или $[f'(x)]' = f''(x)$. Дифференциал, взятый от дифференциала $dy = y'_x\Delta x$, называется *дифференциалом второго порядка*. Этот дифференциал обозначается символом $d(dy) = d^2y$ или $d[df(x)] = d^2f(x)$.

Можно показать, что дифференциал второго порядка равен

$$d^2y = y''_{xx}\,dx^2, \qquad (3.21)$$

а вторая производная

$$y''_{xx} = \frac{d^2y}{d^2x}. \qquad (3.22)$$

Последовательно повторяя описанные операции, можно получить производные третьего, четвертого, ..., n-го порядков и соответственно находить выражения для третьего, четвертого, ..., n-го дифференциалов.

§ 3. Исследование функций, аппроксимирующих лингвистические процессы

Чтобы оценить математическую модель, описывающую то или иное лингвистическое явление, необходимо исследовать все особенности изменения функции (или, как говорят, ее поведение) при изменении аргумента. Это исследование осуществляется с помощью таких математических понятий, как непрерывность функции, ее возрастание и убывание, экстремальные значения, выпуклость и вогнутость, точки перегиба.

1. Непрерывность и точки разрыва функции. Функция $y = f(x)$ называется *непрерывной* в точке x_0, если выполняются следующие условия:

1) функция $y = f(x)$ существует (определена) при $x = x_0$;

2) предел приращения Δy функции y равен нулю, если $\Delta x \to 0$ при $x \to x_0$, т. е.

$$\lim_{\Delta x \to 0} \Delta y = \lim_{\Delta x \to 0} [f(x_0 + \Delta x) - f(x_0)] = 0. \qquad (3.23)$$

Функция, непрерывная во всех точках некоторого отрезка, называется *непрерывной на* этом *отрезке*, а непрерывная во всей области существования — *непрерывной* функцией (примерами непрерывных функций служат линейная и показательная функции или функции $y = \sin x$ и $y = \cos x$; см. гл. 1, § 6).

Если при $x = x_0$ условия непрерывности не выполняются, то точка x_0 называется *точкой разрыва*, а сама функция — *разрывной* в этой точке. Примерами таких функций являются гиперболическая зависимость $y = a/x$, которая терпит разрыв в точке $x = 0$ (см. рис. 9), а также тригонометрические функции $y = \operatorname{tg} x$ и $y = \operatorname{ctg} x$, являющиеся разрывными соответственно в точках $x = (2k + 1)\pi/2$ и $x = k\pi$; где $k = 0, \pm 1, \pm 2, \ldots$ (см. рис. 13).

2. Возрастание и убывание функции. Фунция $y = f(x)$ называется *возрастающей* в интервале (m, n), если большему значению аргумента в этом интервале соответствует большее значение функции (см. рис. 10). Если же в некотором интервале большему значению аргумента соответствует меньшее значение функции, то функция называется *убывающей* в этом интервале (см. рис. 11). Как возрастающие, так и убывающие функции называются *монотонными*, а интервалы возрастания и убывания функции — *интервалами монотонности*.

Исследование функций опирается на простую связь, существующую между поведением функции в некотором интервале и свойствами ее производной в этом же интервале. Эта связь описывается с помощью трех теорем.

Теорема 1 (достаточный признак возрастания функции). *Если производная функции $y = f(x)$ положительна для всех значений x в интервале (m, n), то функция в этом интервале возрастает.*

Теорема 2 (достаточный признак убывания функции). *Если производная функции $y = f(x)$ отрицательна для всех значений x в интервале (m, n), то функция в этом интервале убывает.*

Теорема 3 (достаточный признак постоянства функции). *Если производная функции $y = f(x)$ равна нулю для всех значений x в интервале (m, n), то функция в этом интервале не изменяется, т. е. является постоянной.*

Иногда эти теоремы объединяются под названием достаточного признака монотонности функции.

3. Экстремальные значения функций. Рассмотрим еще раз часть графика функции

$$I = r_0 + \sum_{k=1}^{5} r_k \sin(2\pi k x/l + \varphi_k),$$

аппроксимирующей распределение информации в 12-буквенном русском слове (см. рис. 25).

Наблюдая за изменением функции, нетрудно заметить, что ординаты кривой возрастают на участках BC и DE, достигая в точках C и E максимальной величины по сравнению с ординатами соседних с ними точек. Значение аргумента, при котором функция имеет наибольшую величину, называется *точкой максимума*, а соответствующее значение функции — *максимумом* функции (точки C, E).

Таким образом, функция $y = f(x)$ имеет максимум в точке $x = a$, если при всех x, достаточно близких к a, выполняется неравенство $f(a) > f(x)$.

На участках AB и CD ординаты точек последовательно убывают, принимая наименьшие значения в точках B и D. Значение аргумента, при котором функция принимает наименьшую величину, называется *точкой минимума*, а соответствующее значение функции — *минимумом* функции.

Иными словами, функция $y = f(x)$ имеет минимум в точке $x = a$, если для всех x, достаточно близких к a, выполняется неравенство $f(a) < f(x)$.

Нетрудно заметить, что максимум является границей перехода от возрастания функции к ее убыванию, а минимум — от убывания к возрастанию функции. Понятия «максимум» и «минимум» объединяются обычно одним термином — *экстремум* (или *экстремальное значение*) функции. Понятие экстремума выступает в качестве локального свойства функции, характеризующего ее поведение лишь в непосредственной близости от данной точки $x = a$.

При исследовании функций их экстремальные значения определяются с помощью следующих признаков.

Необходимый признак экстремума формулируется таким образом: *если функция $y = f(x)$ имеет экстремум*

в точке $x = a$, то ее производная в этой точке либо равна нулю [$f'(a) = 0$], *либо вообще не существует.* Этот признак является необходимым, но не достаточным признаком экстремума: из того, что производная в данной точке обращается в нуль или вовсе не существует, еще не следует, что эта точка обязательно будет точкой экстремума. Поэтому при выявлении экстремальных значений приходится оперировать двумя более сильными — так называемыми достаточными признаками экстремума.

Первый достаточный признак экстремума формулируется так: *точка $x = a$ служит точкой экстремума функции $y=f(x)$, если производная $f'(x)$ при переходе x через a меняет знак; при перемене знака с «+» на «—» точка a является точкой максимума, при перемене знака с «—» на «+» — точкой минимума* (в самой точке a производная либо равна нулю, либо не существует).

Исходя из необходимого и первого достаточного признака экстремума, получаем следующее правило исследования функции $y = f(x)$ на максимум и минимум с помощью первой производной:

1) найти первую производную $y'_x = f'(x)$;

2) найти точки, в которых $f'(x) = 0$ или $f'(x)$ не существует; эти точки называются *критическими точками I рода* (если таких точек не существует, то функция экстремумов не имеет);

3) исследовать изменение знака производной при переходе слева направо через каждую критическую точку $x=a$, для чего следует дать x значение несколько меньшее, чем a (т. е. $x = a - h$), а затем несколько большее, чем a (т. е. $x = a + h$), причем под h подразумевается произвольное достаточно малое положительное число (возникающие здесь варианты представлены в табл. 3.6);

4) вычислить экстремальные значения функции в точке a.

Таблица 3.6

$f'(a-h)$	$f'(a)$	$f'(a+h)$	Характер критической точки
+	0 или ∞	−	максимум
−	0 или ∞	+	минимум
+	0 или ∞	+	экстремума нет, функция возрастает
−	0 или ∞	−	экстремума нет, функция убывает

Иногда для установления максимумов и минимумов функции бывает удобнее и проще воспользоваться вторым достаточным признаком экстремума, который формулируется так: *точка $x = a$ есть точка экстремума функции $y = f(x)$, если $f'(a) = 0$, а $f''(a) \neq 0$, причем в том случае, когда $f''(a) > 0$, точка a является точкой минимума, а когда $f''(a) < 0$ — точкой максимума.*

Исходя из второго достаточного признака сформулируем следующее правило исследования функции на экстремум:

1) найти первую и вторую производные;

2) найти критические точки функции;

3) определить знак второй производной: если вторая производная в критической точке положительна, то эта точка является точкой минимума; если вторая производная отрицательна, то рассматриваемая точка есть точка максимума; если же вторая производная в критической точке равна нулю, то следует возвратиться к исследованию функции с помощью первой производной (см. табл. 3.7);

4) вычислить экстремальные значения функции в точке a.

Таблица 3.7

x	$f'(x)$	$f''(x)$	Характер критической точки
a	0	$f''(x) < 0$ $f''(x) > 0$ $f''(x) = 0$	максимум минимум неизвестен (правило неприменимо)

4. Выпуклость, вогнутость и точка перегиба кривой. Для определения понятий выпуклости и вогнутости рассмотрим кривую $y = f(x)$, изображенную на рис. 31. В промежутке (a, b) выберем несколько точек и проведем в них касательные. Из графика видно, что все точки кривой в промежутке (a, b) лежат ниже любой ее касательной в этом промежутке. В таком случае говорят, что кривая *выпукла вверх* в интервале (a, b) или просто *выпукла*.

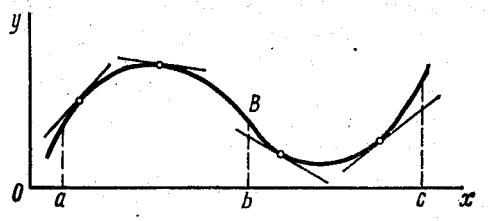


Рис. 31

Рассмотрим теперь другой промежуток (b, c) той же кривой и проведем несколько касательных в точках этого промежутка. Здесь все точки кривой лежат выше любой ее касательной в промежутке (b, c). В таком случае говорят, что кривая *выпукла вниз* в интервале (b, c) или просто *вогнута*.

Вогнутость и выпуклость кривой может быть определена с помощью одного из следующих признаков.

Признак 1. *Интервалу убывания первой производной $f'(x)$ соответствует участок выпуклости (a, b) кривой, а интервалу возрастания $f'(x)$ — участок вогнутости (b, c).*

Признак 2. *Если вторая производная функции $y = f(x)$ во всех точках интервала (a, b) отрицательна, т. е. $f''(x) < 0$, то кривая $y = f(x)$ выпукла в этом интервале и, наоборот, если вторая производная во всех точках интервала (a, b) положительна, т. е. $f''(x) > 0$, то кривая $y = f(x)$ вогнута в указанном интервале.*

Если кривая $y = f(x)$ имеет участки выпуклости и вогнутости, то между ними, очевидно, будет существовать какая-то точка, которая не обладает свойствами соседних промежутков. Такая точка называется *точкой перегиба* (точка B на рис. 31). Поскольку по обе стороны от точки перегиба направления выпуклости различны, знаки второй производной $f''(x)$ в соседних точках также различны, а сама производная должна быть равна нулю.

Для нахождения точек перегиба, а также участков выпуклости и вогнутости кривой $y = f(x)$ рекомендуется пользоваться следующим п р а в и л о м:

1) найти вторую производную $y'' = f''(x)$;

2) найти точки, в которых $f''(x) = 0$ или $f''(x)$ не существует, — *критические точки* II *рода*;

3) исследовать изменение знака второй производной при переходе слева направо через каждую критическую точку II рода $x = b$, для чего следует определить знаки $f''(x)$ в точках $x = b - h$ и $x = b + h$, где h — произвольное достаточно малое положительное число; если $f''(x)$ меняет знак при переходе через точку $x = b$, то график функции имеет точку перегиба; если же знак $f''(x)$ не меняется, то точки перегиба нет — при $f''(x) < 0$ кривая в рассматриваемом интервале выпукла, а при $f''(x) > 0$ — вогнута;

4) вычислить ординаты точек перегиба.

5. Асимптоты. Прямая линия A называется *асимптотой* кривой, если расстояние точки M от прямой A стремится к нулю при неограниченном удалении этой точки по кривой от начала координат. Примером г о р и з о н т а л ь н о й асимптоты может служить ось Ox на рис. 9, примером в е р т и к а л ь н о й асимптоты является ось Oy на том же рисунке. Кроме того, могут существовать и н а к л о н н ы е асимптоты.

Для выявления вертикальных асимптот необходимо найти точки бесконечных разрывов, т. е. точки, в которых $f(x) \to \infty$. Взаимное расположение бесконечной ветви кривой и асимптоты определяется в ходе исследования «знака бесконечности» при стремлении x к x_0 слева и справа. Так, например, для функции $y = \operatorname{tg} x$ значения $\operatorname{tg} x \to +\infty$, когда $x \to \pi/2$, оставаясь меньше чем $\pi/2$, и $\operatorname{tg} x \to -\infty$, когда $x \to \pi/2$, оставаясь больше чем $\pi/2$ (см. рис. 13). Если же $f(x)$ ни при каком значении x не стремится к бесконечности, то вертикальных асимптот нет.

Чтобы найти наклонную асимптоту кривой $y = f(x)$, нужно выразить расстояние между точкой $M(x; y)$ и точкой $M'(x; y')$ прямой A через разность ординат этих точек при одной и той же абсциссе x:

$$y - y' = f(x) - (kx + b).$$

По определению, прямая $y' = kx + b$ служит асимптотой кривой $y = f(x)$ в том случае, если расстояние между ними (разность ординат $y - y'$) стремится к нулю при бесконечном возрастании x, т. е.

$$\lim_{x \to \infty} [f(x) - (kx + b)] = 0. \qquad (3.24)$$

Определим величины k и b. Для нахождения k представим выражение (3.24) в следующем виде:

$$\lim_{x \to \infty} x \left[\frac{f(x)}{x} - k - \frac{b}{x} \right] = 0.$$

Так как первый множитель этого произведения стремится к бесконечности, то второй множитель должен стремиться к нулю. Отсюда

$$\lim_{x \to \infty} \left[\frac{f(x)}{x} - k - \frac{b}{x} \right] = 0,$$

или

$$\lim_{x \to \infty} \frac{f(x)}{x} - \lim_{x \to \infty} \frac{b}{x} = k,$$

но $\lim\limits_{x \to \infty} \frac{b}{x} = 0$, поэтому

$$\lim_{x \to \infty} \frac{f(x)}{x} = k. \qquad (3.25)$$

Найдя величину k, теперь из равенства (3.24) определим b:

$$b = \lim_{x \to \infty} [f(x) - kx]. \qquad (3.26)$$

Для установления взаимного расположения кривой и асимптоты следует выяснить знак разности $f(x) - (kx + b)$ как при $x \to +\infty$, так и при $x \to -\infty$. Если эта разность положительна, то кривая расположена н а д асимптотой; если же она отрицательна, то кривая лежит п о д асимптотой.

6. Общая схема исследования функции и построение ее графика. Как уже говорилось, при математической экспликации того или иного лингвистического процесса представляют интерес не отдельные численные значения моделирующей его функции, а существенные особенности этой функции, например: возрастание и убывание, максимум и минимум, выпуклость и вогнутость ее графика. Все эти свойства функции хорошо прослеживаются на графике. Однако построение графика, так же как и выбор аналитической формы функции по некоторому (пусть даже очень большому) количеству точек, взятых наудачу, всегда связано с опасностью упустить некоторые существенные особенности функции. Напротив, предварительное аналитическое исследование, выявляя ее существенные особенности, позволяет определить положение характерных точек.

Исследование функции, а затем построение графика осуществляется обычно по схеме, включающей следующие этапы.

1. Определение области существования и промежутков непрерывности функции.

2. Установление области возрастания и убывания функции; нахождение точек экстремума и выяснение их характера.

3. Установление промежутков выпуклости и вогнутости кривой; нахождение ее точек перегиба.

4. Отыскание асимптот.

5. Сведение результатов исследования в таблицу.

Эта таблица должна включать значения функции в ее характерных точках, к которым относятся точки экстремума, точки перегиба и точки пересечения функции с осями координат, точки разрыва (если они существуют), а также границы области существования функции. Кроме того, в таблицу вносятся и некоторые дополнительные значения функции. Перенеся найденные точки на чертеж и соединив их плавной кривой, мы получим график исследуемой лингвистической функции.

Проведя исследование функции и представив его результаты в легко обозримой графической форме, мы получаем возможность объективно решить вопрос о соответствии математической модели ее лингвистическому оригиналу.

7. Исследование двух функций, аппроксимирующих распределение информации в тексте. Моделирование распределения информации в тексте осуществляется с помощью двух функций: выражения

$$I_n^T = H_1\left[1 - a\left(1 - \frac{1}{n}\right)\right], \qquad (3.27)$$

где n — номер буквы в тексте, H_1 — энтропия первой буквы текста, a — коэффициент [23, с. 67], а также показательной функции

$$I_n = (I_0 - I_\infty)\, e^{-sn} + I_\infty. \qquad (3.28)$$

Необходимо выяснить, какая из этих зависимостей лучше моделирует распределение информации в тексте. Чтобы сделать возможным использование аппарата исследования функций, будем считать аргумент n в обоих случаях непрерывной величиной.

1) Приведем правую часть соотношения (3.27) к общему знаменателю:

$$I_n^T = \frac{H_1 n - aH_1 n + aH_1}{n}.$$

Из полученного равенства видно, что функция $I_n^T = y = f(n)$ существует и непрерывна при всех значениях аргумента n, кроме $n = 0$, т. е. в интервалах $(-\infty, 0)$, $(0, +\infty)$.

2) Установим теперь характер поведения функции в интервале $(0, +\infty)$. Для этого найдем первую производную

$$f'(n) = \left(H_1 - H_1 a + \frac{H_1 a}{n}\right)'_n = -H_1 a \frac{1}{n^2}. \qquad (3.29)$$

Это соотношение показывает, что при любом значении аргумента n в интервале $(0, +\infty)$ первая производная является отрицательной величиной, откуда следует, что данная функция убывает в указанном интервале.

Так как среди возможных значений аргумента нельзя найти такое, при котором производная $f'(n)$ обращалась бы в нуль, то функция (3.27) не имеет точек экстремума.

3) Для установления характера кривой в интервале $(0, +\infty)$ (выпуклости или вогнутости) найдем вторую производную. Имеем

$$f''(n) = \left(-H_1 a \frac{1}{n^2}\right)'_n = \frac{2H_1 a}{n^3}.$$

Очевидно, что при всех положительных значениях n вторая производная данной функции положительна. Значит, график распределения информации в тексте представляет собой вогнутую линию и не имеет в интервале $(0, +\infty)$ точек перегиба.

4) Определим асимптоты кривой распределения информации. Так как при $n \to 0$

$$\lim_{n\to 0} f(n) = \lim_{n\to 0} \frac{H_1 n - aH_1 n + aH_1}{n} = +\infty,$$

то прямая $n = 0$ (т. е. ось ординат) есть вертикальная асимптота.

Для определения положения наклонной асимптоты нужно воспользоваться формулами (3.25) и (3.26). Угловой коэффициент равен

$$k = \lim_{n\to\infty} \frac{f(n)}{n} = \lim_{n\to\infty} \frac{H_1 n - aH_1 n + aH_1}{n^2} =$$

$$= \lim_{n\to\infty} \left(\frac{H_1}{n} - \frac{aH_1}{n} + \frac{aH_1}{n^2}\right) = 0.$$

Начальная ордината асимптоты

$$b = \lim_{n\to\infty} [f(n) - kn] = \lim_{n\to\infty} \frac{H_1 n - aH_1 n + aH_1}{n} =$$

$$= \lim_{n\to\infty} \left(H_1 - aH_1 + \frac{aH_1}{n}\right) = H_1(1 - a).$$

Поскольку угловой коэффициент асимптоты равен нулю, эта асимптота имеет горизонтальное положение, а ее уравнение принимает вид

$$y = H_1(1 - a). \qquad (3.30)$$

Для установления взаимного расположения кривой распределения информации и горизонтальной асимптоты определим разность ординат этой кривой и асимптоты при одном и том же значении n:

$$\left(H_1 - aH_1 + \frac{aH_1}{n}\right) - (H_1 - aH_1) = \frac{aH_1}{n}.$$

Так как полученная разность положительна (величины a, H_1 и n положительны), то горизонтальная асимптота (3.30) расположена под кривой (3.27).

Перенеся все результаты исследования на чертеж, мы получаем график функции (3.27) в том виде, как он изображен на рис. 32.

Используя ту же схему, исследуем функцию (3.28), представляющую собой второй вариант описания распределения информации в тексте. Здесь мы получаем следующие результаты.

1) Функция $I_n = y = \varphi(n)$ существует и непрерывна на всей числовой оси: $-\infty < n < +\infty$.

2) Вид первой производной

$$\varphi'(n) = -s(I_0 - I_\infty)e^{-sn}$$

показывает, что функция является убывающей, так как первая производная всегда отрицательна, и экстремальных значений не имеет.

3) Вторая производная

$$\varphi''(n) = s^2(I_0 - I_\infty)e^{-sn}$$

положительна, поэтому кривая (3.28) вогнута и не имеет точек перегиба.

4) График функции имеет только одну горизонтальную асимптоту $y = I_\infty$. При $n = 0$ кривая пересекает ось Oy в точке $y = I_0$.

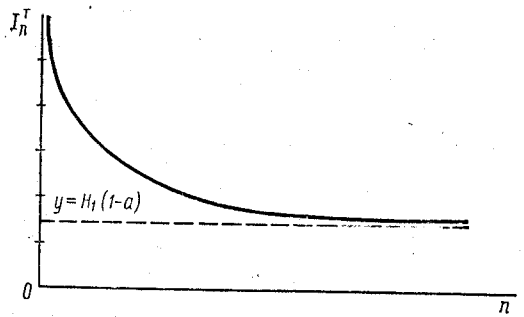


Рис. 32

Всем этим условиям соответствует уже знакомый нам график, изображенный на рис. 19.

Результаты исследования показывают, что зависимость

$$I_n^T = H_1\left[1 - a\left(1 - \frac{1}{n}\right)\right]$$

моделирует распределение информации в тексте при $0 < n < \infty$, в то время как зависимость

$$I_n = (I_0 - I_\infty)e^{-sn} + I_\infty$$

описывает распределение информации при $0 \leqslant n < \infty$.

При информационных измерениях языка широко используется величина I_0 (информация алфавита). Эта величина характеризует количество информации (неопределенности), возникающее в том случае, когда совершенно не учитываются дистрибутивно-вероятностные ограничения на употребление лингвистических единиц, причем n здесь считается равным нулю. Эта ситуация учтена в зависимости (3.28), но не предусмотрена в функции (3.27), которая при $n = 0$ не существует.

Из всего сказанного следует, что для моделирования распределения информации в тексте следует выбрать функцию (3.28).

ГЛАВА 4

СУММИРОВАНИЕ И ИНТЕГРИРОВАНИЕ В ЛИНГВИСТИЧЕСКИХ ПРОЦЕССАХ

Многие задачи квантитативной лингвистики возникают как задачи вычисления сумм языковых величин, характеризующих тот или иной речевой или диахронический процесс в целом или в его части. Эти задачи решаются с помощью аппарата теории рядов и интегрирования.

§ 1. Основные понятия теории рядов

1. Числовой ряд. Если взять множество значений величин информации, приходящихся на первую, вторую и т. д. буквы слова или множество относительных или абсолютных частот, характеризующих употребительность данного лингвистического явления в определенные периоды истории языка, то видно, что из этих множеств можно образовать числовые последовательности, подчиняющиеся определенному правилу. *Бесконечная числовая последовательность* имеет следующий вид:

$$u_1, u_2, \ldots, u_n, \ldots,$$

где u_n — общий член последовательности.

Члены последовательности можно суммировать. Тогда алгебраическая сумма членов последовательности

$$u_1 + u_2 + u_3 + \ldots + u_n + \ldots \qquad (4.1)$$

образует *бесконечный числовой ряд*, или просто *ряд*. Иногда для обозначения ряда применяют запись $\sum_{n=1}^{\infty} u_n$. Выражение для n-го члена ряда при произвольном n называется *общим членом* ряда. Обычно общий член ряда задается формулой $u_n = f(n)$, пользуясь которой, можно сразу написать любой член ряда. Например, в ряде, характеризующем сумму информаций, приходящихся на первую, вторую, ..., n-ю, ... буквы слова, общий член имеет вид

$$u_n = I_n = I_0 e^{-sn}.$$

Составим из первых членов ряда (4.1) суммы

$$S_1 = u_1,$$
$$S_2 = u_1 + u_2,$$
$$S_3 = u_1 + u_2 + u_3,$$
$$\ldots\ldots\ldots\ldots$$
$$S_n = u_1 + u_2 + u_3 + \ldots + u_n,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$

которые называются *частичными суммами* данного ряда.

Если при бесконечном возрастании номера n частичная сумма ряда S_n стремится к конечному пределу S, т. е.

$$\lim_{n\to\infty} S_n = S, \tag{4.2}$$

то в этом случае ряд называется *сходящимся*, а число S — его *суммой*.

Если же величина S_n неограниченно возрастает, т. е. $\lim\limits_{n\to\infty} S_n = \infty$, или S_n хотя и не возрастает неограниченно, но ни к какому конкретному пределу не стремится, то мы имеем дело с *расходящимся рядом*.

Ряды могут быть либо *знакоположительными* (все члены такого ряда положительны, т. е. $u_n > 0$), либо *знакоотрицательными* (все члены такого ряда отрицательны, т. е. $u_n < 0$), либо *знакопеременными* (члены такого ряда могут быть как положительными, так и отрицательными).

Членами ряда могут служить не только числа, но и функции некоторого аргумента x. В этом случае мы имеем дело с так называемым *функциональным рядом*, который имеет вид

$$f_1(x) + f_2(x) + f_3(x) + \ldots + f_n(x) + \ldots, \tag{4.3}$$

где функции $f_1(x)$, $f_2(x)$, ..., $f_n(x)$, ... определены в некоторой области изменения аргумента x. Придавая x какое-либо значение x_0 из области определения функций $f_n(x)$, получим числовой ряд

$$f_1(x_0) + f_2(x_0) + \ldots + f_n(x_0) + \ldots .$$

Этот ряд может сходиться или расходиться. Если он сходится, то точка x_0 называется *точкой сходимости* функционального ряда (4.3). Совокупность всех точек сходимости функционально о ряда называется *областью* его *сходимости*.

Важным случаем функциональных рядов являются *степенные ряды*. Степенной ряд имеет вид

$$a_0 + a_1x + a_2x^2 + \ldots + a_nx^n + \ldots, \tag{4.4}$$

где x — аргумент, а a_0, a_1, a_2, ..., a_n — постоянные числа, называемые коэффициентами.

Степенной ряд можно почленно дифференцировать сколько угодно раз, причем полученные при этом ряды будут иметь ту же область сходимости.

Пусть функция $f(x)$ является суммой степенного ряда:

$$f(x) = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n + \ldots . \tag{4.5}$$

Последовательно дифференцируя n раз степенной ряд (4.5) и полагая в полученных равенствах $x = 0$, найдем следующие значения коэффициентов:

$$a_0 = f(0),\quad a_1 = f'(0),\quad a_2 = \frac{f''(0)}{1\cdot 2},$$

$$a_3 = \frac{f'''(0)}{1\cdot 2\cdot 3},\ \ldots,\quad a_n = \frac{f^{(n)}(0)}{n!}.$$

Подставив значения коэффициентов в степенной ряд (4.5), получим ряд

$$f(x) = f(0) + f'(0)\,x + \frac{f''(0)}{1\cdot 2}x^2 + \frac{f'''(0)}{1\cdot 2\cdot 3}x^3 + \ldots + \frac{f^{(n)}(0)}{n!}x^n + \ldots, \tag{4.6}$$

который называется *рядом Маклорена*. С помощью ряда Маклорена любую функцию $f(x)$ можно разложить в ряд по степеням x.

Например, разложение для функции $f(x) = e^x$ получается следующим образом. Сначала определим значения производных $f'(x) = e^x$, $f''(x) = e^x$, $f'''(x) = e^x$, ... Полагая $x = 0$, найдем $f(0) = 1$, $f'(0) = 1$, $f''(0) = 1$, $f'''(0) = 1$, ... Подставив значения производных в ряд Маклорена, имеем

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!} + \ldots . \tag{4.7}$$

Если в равенстве (4.7) положить $x = 1$, то получим значение «числа Эйлера» в виде суммы числового ряда:

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \ldots + \frac{1}{n!} + \ldots .$$

Представление числа e в виде суммы числового ряда дает возможность вычислять эту величину с любой степенью точности.

2. Признаки сходимости ряда. Важнейший вопрос исследования ряда состоит в определении его сходимости или расходимости. Поэтому рассмотрим те необходимые и достаточные признаки, с помощью которых по общему члену ряда можно определять его сходимость или расходимость.

Необходимый признак сходимости ряда формулируется следующим образом: *если ряд*

$$u_1 + u_2 + \ldots + u_n + \ldots$$

сходится, то необходимо, чтобы общий член ряда u_n при неограниченном возрастании n стремился к нулю:

$$\lim_{n\to\infty} u_n = 0.$$

Из необходимого условия сходимости ряда вытекает достаточный признак расходимости ряда: *если общий член ряда u_n не стремится к нулю, то ряд является расходящимся.*

Из достаточных признаков сходимости ряда мы будем применять следующие три признака.

Первый признак сходимости (**признак сравнения**). *Пусть даны два знакоположительных ряда:*

$$\sum_{n=1}^{\infty} u_n = u_1 + u_2 + u_3 + \ldots + u_n + \ldots \ (u_n > 0) \tag{4.8}$$

и

$$\sum_{n=1}^{\infty} v_n = v_1 + v_2 + v_3 + \ldots + v_n + \ldots \ (v_n > 0), \tag{4.9}$$

причем *каждый член ряда* (4.8) *не больше соответствующего члена ряда* (4.9), т. е. $u_n \leqslant v_n$ $(n = 1, 2, \ldots)$. *Тогда если сходится ряд* (4.9), *то сходится и ряд* (4.8).

Легко показать, что из этого достаточного признака сходимости ряда вытекает достаточный признак расходимости: *если ряд* (4.8) *расходится, то расходится и ряд* (4.9).

При использовании только что описанного признака об одном из сравниваемых рядов должно быть заранее известно, что он является сходящимся или расходящимся.

Чаще всего в качестве эталона сходящегося ряда принимается *геометрическая прогрессия*

$$+ a_1q + a_1q^2 + a_1q^3 + \ldots + a_1q^{n-1} + \ldots . \qquad (4.10)$$

при условии, что знаменатель $|q| < 1$.

Эталоном расходящегося ряда часто служит *гармонический ряд*

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots + \frac{1}{n} + \ldots . \qquad (4.11)$$

В лингвистической практике часто встречаются такие ряды, которые невозможно сопоставить с каким-либо уже изученным с точки зрения признака сходимости рядом. В этом случае приходится использовать другие достаточные признаки сходимости.

Второй признак сходимости (**признак Даламбера**). *Если для знакоположительного ряда*

$$u_1 + u_2 + \ldots + u_n + u_{n+1} + \ldots \qquad (4.12)$$

выполняется условие

$$\lim_{n\to\infty} \frac{u_{n+1}}{u_n} = k$$

то при $k < 1$ *ряд сходится, при* $k > 1$ *— расходится, а при* $k = 1$ *вопрос о сходимости ряда неопределен: в одних случаях ряд сходится, в других — расходится.*

Третий признак сходимости (**признак Коши**). *Если для знакоположительного ряда* (4.12) *выполняется условие*

$$\lim_{n\to\infty} \sqrt[n]{u_n} = k,$$

то при $k < 1$ *ряд сходится, при* $k > 1$ *— расходится, при* $k = 1$ *вопрос о сходимости ряда остается неопределенным.*

3. Абсолютно сходящиеся ряды. До сих пор мы говорили о признаках сходимости знакоположительных рядов. Этими признаками можно воспользоваться при определении условий сходимости знакопеременных рядов.

Пусть дан ряд, содержащий как положительные, так и отрицательные члены:

$$u_1 + u_2 + u_3 + \ldots + u_n + \ldots . \qquad (4.13)$$

Очевидно, что такой ряд можно сделать знакоположительным. Так, заменив члены ряда (4.13) их абсолютными величинами, получим

$$|u_1| + |u_2| + |u_3| + \ldots + |u_n| + \ldots . \qquad (4.14)$$

Поскольку сумма членов знакопеременного ряда всегда меньше суммы абсолютных величин этих членов, можно утверждать, что *знакопеременный ряд* (4.13) *сходится, если сходится ряд* (4.14), *составленный из абсолютных величин его членов.* В описанном случае сходящийся знакопеременный ряд называют *абсолютно сходящимся.*

§ 2. Каков максимальный объем информации в слове?

1. Можно ли рассматривать слово в качестве ряда? Прежде чем ставить задачу, связанную с оценкой той максимальной информации, которую может нести слово, необходимо решить вопрос о том, вправе ли мы рассматривать связный текст и слово в качестве ряда и что это будет за ряд. Что касается текста, то здесь особых трудностей не возникает: каждый текст практически может быть продолжен до бесконечности (см. гл. 1, § 8, п. 3). и существует функция $I_n = f(n)$, позволяющая определять количество информации, приходящееся на каждую его букву при условии, что $n \to \infty$.

Сложнее обстоит дело со словом. Действительно, может ли слово быть представлено в виде бесконечного ряда величин информации? Ведь каждое слово состоит обычно из конечного и сравнительно небольшого числа дискретных единиц (букв, слогов, морфем, фонем), и трудно себе представить слово какого-либо конкретного языка, состоящее, например, из двухсот букв.

Разумеется, каждая конкретная словоформа представляет собой конечную последовательность букв или фонем, обычно не сильно отклоняющуюся от средней длины слова, характеризующей данный язык (об этом будет подробнее сказано в гл.7).

Вместе с тем, даже если следовать критерию графической цельнооформленности, трудно указать конкретный предел длины словоформы. Это касается не только языка вообще, но и конкретных языков. Даже в тех индоевропейских языках, в которых флективно-аналитическая морфология и слабые возможности словообразования и словосложения накладывают ограничения на образование слишком длинных слов, мы постоянно встречаемся с такими многобуквенными словоформами, как англ. transsubstantiation' возможность переноса доказательства', или нем. Kesselsteinverhinderungsmittelerzeugungsgesellschaft', компания по изготовлению растворителей для котельной накипи, или русск. *двустороннесимметричный, информационно-статистический, невосстанавливаемость.* При этом для каждой длинной словоформы можно теоретически образовать еще более длинную ее производную —ср. transsubstantiation — intrans-

substantiation, *двустороннесимметричный — двустороннесиметричнообразный* и т. п.

Что же касается инкорпорирующих и агглютинирующих языков*, то сказать, на каком максимальном удалении от начала слова должен находиться его конец, в этом случае еще труднее. Ср. в этом смысле такие цельнооформленные образования, как кабард. шыжъыфIытхъуэр' старая хорошая буланная лошадь', чукот. мытрэлгитэкупрэныскивыскычетгъэ'очень стремительно[направимся изготавливать сети', турецк. türkläştirämädiklärimizdänmisiniz' не из тех ли вы, кого мы не могли отуречить', от которых также могут быть образованы еще более длинные производные слова. Так, например, приведенная турецкая словоформа теоретически может быть удлинена путем циклического повторения цепочки аффиксов läştirämädik (правда, такое удлинение допустимо с грамматической, но не смысловой точки зрения)**.

Отсюда можно предположить, что длина орфографического слова (во всяком случае в языках определенного типа) является потенциально бесконечной величиной.

Это предположение станет еще более очевидным, если отнести к словам стилевые образования типа «Нехочунебунехочунебуду» (газ. «Известия», 29. III 73, с. 6).

Теперь попробуем отказаться [62, с. 89] от критерия орфографической цельнооформленности и будем исходить из принципа семантическо-грамматической цельности. В этом случае в категорию слов попадут такие цепочки орфографических слов, как фр. je ne le lui ai pas expliquè'я это ему не объяснил'; англ. (the) Prime Minister of Britain's (residence)' (резиденция) премьер-министра Англии', где присоединение форманта 's к словоформе Englang вм. Minister указывает на то, что цепочка Prime Minister of Britain's воспринимается как слово; русск. *кое в чем* и т. п. При этом подходе словами должны считаться и китайские «цы» — комбинации иероглифов, каждый из которых изображает слог-слово [29, с. 115—116].

Отказ от принципа графической цельнооформленности и включение в состав слов аналитических цепочек и образований типа китайских «цы» делает еще менее фиксированной правую границу слова.

Все это позволяет предположить, что длина абстрактного слова-модели является величиной потенциально бесконечной и образует ряд вычислимых величин информаций

$$I_1 + I_2 + \ldots + I_n + \ldots . \qquad (4.15)$$

* В инкорпорирующих языках (в первую очередь, палеоазиатских, некоторых иберо-кавказских и индейских) широко применяется сочетание примыкающих друг к другу корней, совокупность которых оформляется служебными элементами. Этим путем образуются слова или особого рода синтагмы (ср. приводимые в тексте кабардинский и чукотский примеры). Агглютинирующие языки (в первую очередь, тюркские и финноугорские) широко используют присоединение стандартных аффиксов к неизменяемым основам. Примером агглютинативного слова является приводимый в тексте турецкий пример, в котором к основе türk-присоединена цепочка аффиксов [5, с. 31, 177].

** Турецкий пример сообщен нам А. М. Щербаком.

2. Оценка максимального объема информации, содержащегося в слове. Прежде чем перейти к вычислению общей суммы информации, содержащейся в ряде (4.15), необходимо выяснить, является ли этот ряд сходящимся.

Проверим, выполняется ли необходимое условие сходимости ряда (4.15). Так как его общий член

$$I_n = I_0 e^{-sn} = I_0/e^{sn}$$

при неограниченном возрастании n стремится к нулю, то необходимое условие сходимости выполнено. Воспользовавшись численными значениями параметров для русского слова, взятого вне текста ($I_0 = 5$, $s = 0{,}25$) и в тексте ($I_0 = 5$, $s = 0{,}40$), применим для проверки сходимости ряда достаточный признак Коши. Тогда получаем для обоих случаев:

$$\lim_{n\to\infty} \sqrt[n]{I_0 e^{-sn}} = \lim_{n\to\infty} \sqrt[n]{5e^{-0{,}25n}} = 0{,}7788 < 1$$

и

$$\lim_{n\to\infty} \sqrt[n]{5e^{-0{,}40n}} = 0{,}6703 < 1.$$

Как в одном, так и в другом случае ряды являются сходящимися и можно вычислить их сумму. Чтобы получить теоретическую оценку того количества информации, которое может нести слово максимальной длины, взятое вне текста и в контексте, представим предел частичной суммы ряда (4.15) в следующем виде:

$$\lim_{n\to\infty} S_n = I_0 [(1/e^s)^0 + (1/e^s)^1 + (1/e^s)^2 + \ldots + (1/e^s)^n + \ldots]. \qquad (4.16)$$

Сумма в квадратных скобках представляет собой сумму членов бесконечно убывающей геометрической прогрессии, знаменатель которой равен $1/e^s$. Следовательно, эта сумма равна

$$\lim_{n\to\infty} \frac{1-(1/e^s)^n}{1-1/e^s} = \frac{1}{1-1/e^s}.$$

Подставляя последнее выражение в (4.16), получаем формулу, оценивающую максимальное количество информации, которое может содержаться в слове:

$$I_{\max} = \lim_{n\to\infty} S_n = \frac{I_0}{1-1/e^s}. \qquad (4.17)$$

Заменив коэффициент s численными значениями, получим теоретические оценки максимального количества синтактической информации, которое может быть передано словом, взятым в контексте и вне контекста. Эти оценки относительно некоторых европейских языков приведены в табл. 4.1.

Таблица 4.1

Языки	I_{max} вне текста (дв. ед.)	I_{max} в тексте (дв. ед.)	Языки	I_{max} вне текста (дв. ед.)	I_{max} в тексте (дв. ед.)
Русский	22,5	15,0	Немецкий	21,4	14,3
Болгарский	22,0	14,8	Французск.	21,2	14,1
Английский	21,1	14,1	Румынский	21,6	14,4

Данные табл. 4.1 говорят о том, что максимальные информационные нагрузки текстового и словарного слова в индоевропейских языках примерно одинаковы. Возникает вопрос, сохраняют ли эти количественные оценки свою силу для агглютинирующих и инкорпорирующих языков? Если да, то мы, очевидно, имеем дело с квантитативными свойствами оперативной памяти человека, если нет, то речь идет о новых квантитативных признаках типологии языков.

Что касается применения аппарата теории рядов к вычислению информации, содержащейся в тексте, то этот аппарат может быть применен к участкам текста конечной длины. Если же речь идет о тексте неопределенной длины, то он формализуется с помощью ряда

$$I_1 + I_{II} + \ldots + I_n + \ldots, \tag{4.18}$$

общим членом которого служит выражение

$$I_n = (I_0 - I_\infty)\, e^{-sn} + I_\infty. \tag{4.19}$$

При бесконечном возрастании n выражение (4.19) стремится к I_∞ (см. гл. 1, §8, п. 3), что говорит о расходимости ряда. Поэтому сумма членов ряда (4.18) предела не имеет.

3. Ряд и диахронический процесс. Диахронические процессы также могут быть исследованы с помощью рядов. Так, например, при статистическом анализе старо- и средне-французских текстов разных эпох выяснилось, что сумма долей употребления форм именительного падежа единственного числа существительных мужского рода, восходящих к латинским существительным II, III и IV склонения (ср. murus > murs > mur, tempus > temps), для различных эпох истории французского языка может быть представлена в следующем виде:

$$1 + 0{,}25 + 0{,}036 + 0{,}015 + \ldots .$$

Перепишем этот ряд таким образом:

$$1 + \frac{1}{4} + \frac{1}{27} + \frac{1}{64} + \ldots,$$

или

$$\frac{1}{1^1} + \frac{1}{2^2} + \frac{1}{3^3} + \frac{1}{4^4} + \ldots . \tag{4.20}$$

Сравнивая полученный ряд с бесконечно убывающей геометрической прогрессией (4.10), где $a_1 = 1$, $q = 1/2$, мы убеждаемся, что все члены ряда (4.20), начиная со второго, меньше соответствующих членов убывающей геометрической прогрессии. Но ряд (4.10) является сходящимся, поэтому сходится и ряд (4.20). Иными словами, можно ожидать, что в какой-то период развития французского языка именные формы единственного числа мужского рода с окончанием s полностью выйдут из употребления, и мы будем иметь конечную и поэтому вычислимую сумму долей этих форм для всех периодов развития французского языка.

§ 3. Лингвистические задачи, приводящие к понятию интеграла

1. Прогнозирование развития терминологии. Среди задач диахронического суммирования, которые не могут быть решены с помощью аппарата теории рядов, рассмотрим задачу прогнозирования роста научно-технической терминологии.

Для лексикографической практики и особенно при построении действующих систем машинного перевода и реферирования важно иметь прогноз количественного роста терминологии в различных областях знаний. Такое прогнозирование давало бы возможность сознательно планировать выпуск двуязычных и одноязычных политехнических, а также отраслевых словарей и справочников. Этот прогноз позволил бы строить достаточно эффективные алгоритмы пополнения машинных словарей.

Будем строить этот прогноз относительно некоторого идеализированного процесса развития терминологии, при котором прирост новых терминов на протяжении равных отрезков времени последовательно увеличивается.

Задача такого прогнозирования решалась бы просто, если бы все новые слова и выражения, появляющиеся в научно-технической литературе за определенный промежуток времени (например, десятилетие) фиксировались и подсчитывались лексикографической службой в конце этого периода. Тогда в течение каждого десятилетия количество узаконенных в словарях терминов оставалось бы неизменным. Зная число новых радиотехнических, кибернетических, ракетостроительных и т. п. терминов, вошедших в обиход в начале 50-х, 60-х, 70-х годов нашего столетия, мы могли бы предсказать, сколько новых слов и выражений из соответствующих отраслей науки и техники будет введено в научный обиход в начале 80-х, 90-х годов, а также в начале первого, второго и т. д. десятилетий следующего века.

В действительности же развитие терминологии представляет собой непрерывный процесс: статьи и книги, содержащие новые термины, появляются в период всего десятилетия. Поэтому в течение семидесятых годов терминология не остается неизменной по сравнению с 1 января 1970 г., а непрерывно растет. Следовательно, потребность в словарях и справочниках, например в 1977 или 1978 г., выше, чем в 1970 г. Эту потребность нельзя определить ретроспективно и по прогнозу 1980 г., поскольку здесь темп роста терминологии будет заведомо выше уровня 1977-го или 1978-го годов.

Такие затруднения возникают всегда, когда, осуществляя квантитативное описание того или иного диахронического процесса, мы пользуемся лингво-статистическими данными, относящимися к большим хронологическим интервалам (столетие, десятилетие), взятым целиком. Поскольку эти интервалы выступают в виде дискретных отрезков, динамика непрерывного лингвистического процесса внутри этих отрезков ускользает из поля зрения исследователя.

Как же преодолеть это противоречие между «порционным» дискретным характером результатов количественного эксперимента в языкознании и непрерывностью лингвистического процесса?

Очевидно, что точнее описать динамику лингвистического процесса можно было бы с помощью количественных сведений, взятых относительно более мелких хронологических отрезков. Скажем, при прогнозировании развития терминологии вместо десятилетия можно было бы использовать статистику употребления новых терминов за пять лет или за год.

Рассмотрим этот вопрос более подробно.

Пусть в момент x темп роста терминологии составляет $f(x)$ новых лексических единиц в десять лет. Это значит, что если указанный темп развития сохранится, то в течение десятилетия, к которому принадлежит момент x, прирост новых терминов составит $f(x)$. В действительности же через год может появиться фундаментальное исследование в данной области, и скорость развития терминологии $f(x)$ изменится. Чтобы более точно оценить ход нашего процесса, уменьшим интервал и определим с помощью величины $f(x)$ увеличение числа терминов в конце первого года десятилетия до выхода новой книги. Очевидно, что приращение будет равно $\frac{1}{10} f(x)$, поскольку $f(x)$ указывает количество новых терминов в десятилетие.

Но и в течение года также могут произойти события, ускоряющие развитие терминологии: отдельные научные открытия, создание принципиально новых технических устройств и т. п. Поэтому в течение года появится не $\frac{1}{10} f(x)$ терминов, а больше. Это количество можно оценить на основе того темпа роста, который будет достигнут через год. Поскольку эта новая скорость развития терминологии измеряется величиной $f(x + \frac{1}{10})$, количество терминов, появившихся за год, составляло бы $\frac{1}{10} f(x + \frac{1}{10})$.

Однако на самом деле за первый год рассматриваемого десятилетия появится меньше новых терминов: ведь скорость будет достигнута лишь в самом конце года.

Итак, истинное количество новых терминов, которое появилось за рассматриваемый год, остается неизвестным; известно лишь, что оно больше $\frac{1}{10} f(x)$ и меньше $\frac{1}{10} f(x + \frac{1}{10})$.

Еще раз уменьшив величину временно́го интервала, попробуем определить прирост новых терминов за месяц, т. е. в промежутке от

x до $x + \frac{1}{120}$. Здесь снова выясняется, что хотя их число точно определить нельзя, можно утверждать, что оно лежит в интервале между $\frac{1}{120} f(x)$ и $\frac{1}{120} f(x + \frac{1}{120})$. Это последовательное уменьшение интервала показано на рис. 33.

Повторяя все эти рассуждения и считая $f(x)$ непрерывной функцией*, снова уменьшим величину интервала. Обозначим этот интервал через Δx; тогда можно сказать, что в период от x до $x + \Delta x$ количество новых терминов колеблется от $f(x)\,\Delta x$ до $f(x + \Delta x)\,\Delta x$.

Теперь возьмем Δx настолько малым, что в этом промежутке темп появления новых терминов будет отличаться от скорости $f(x)$ на величину, меньшую, чем любая наперед заданная бесконечно

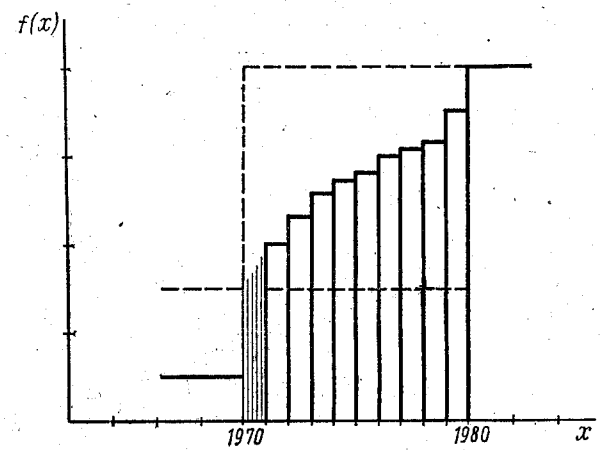


Рис. 33

малая величина α. Иными словами, темп появления новых терминов лежит в промежутке $(f(x) - \alpha, f(x) + \alpha)$. Обозначим прирост терминов за сколь угодно малый промежуток времени Δx через Δy; тогда

$$\Delta y = [f(x) \pm \alpha]\,\Delta x = f(x)\,\Delta x \pm \alpha \Delta x \qquad (4.21)$$

при условии, что $f(x)$ непрерывна в интервале Δx.

Таким образом, если за период Δx приращение новых терминов составляет Δy, то общее количество терминов, накопленных в подъязыке к моменту времени x, равняется величине y.

В правой части равенства (4.21) два слагаемых, из которых первое представляет собой произведение известной функции $f(x)$ на сколь угодно малое приращение аргумента Δx, а второе является

* Само собой разумеется, что число новых терминов, появившихся в определенном временно́м срезе, величина дискретная, однако с точки зрения диахронии ее следует рассматривать как непрерывную величину (см. гл. 1, § 3, п 3).

произведением сколь угодно малой величины Δx на бесконечно малую величину α.

Тогда при условии, что $\Delta x \to 0$, оба слагаемых правой части равенства будут бесконечно малыми величинами. В гл. 2 § 1, п. 3 было показано, что бесконечно малые величины могут обладать разной степенью малости. Выясним, обладают ли одинаковой степенью малости величины $f(x)\Delta x$ и $\alpha\Delta x$. Для этого определим предел отношения этих величин к величине Δx при условии $\Delta x \to 0$:

$$\lim_{\Delta x \to 0} \frac{f(x)\,\Delta x}{\Delta x} = f(x); \quad \lim_{\Delta x \to 0} \frac{\alpha \Delta x}{\Delta x} = 0.$$

Из этого следует, что второе слагаемое $\alpha\Delta x$ является величиной бесконечно малой б о л е е в ы с о к о г о п о р я д к а по сравнению с Δx, и поэтому им можно пренебречь. Что же касается первого слагаемого $f(x)\Delta x$, то оно является г л а в н о й ч а с т ь ю приращения величины y, или ее дифференциалом. Тогда, согласно (3.13), можно записать

$$dy = f(x)\,dx.$$

Отсюда

$$\frac{dy}{dx} = f(x). \tag{4.22}$$

Равенство (4.22) представляет собой известное выражение производной $\frac{dy}{dx} = y'_x$. Поэтому можно сказать, что $f(x)$ есть п р о и з в о д н а я некоторой функции $y = F(x)$.

Итак, если раньше, дифференцируя функцию $y = F(x)$, мы находили ее производную $y'_x = F'(x) = f(x)$ или дифференциал $dy = F'(x)\,dx = f(x)\,dx$, то теперь, как показывает только что рассмотренный пример, возникает обратная задача: для данной функции $f(x)$ требуется найти *первообразную*, т. е. такую функцию $F(x)$, производная которой равнялась бы заданной функции $f(x)$, или, что то же, дифференциал которой равнялся бы заданному выражению $f(x)\,dx$. Операция отыскания первообразной называется *интегрированием*. Если рассматривать дифференцирование в качестве прямого действия, то интегрирование выступает в качестве обратного действия.

Интегрирование может быть применено не только в диахронии, но и при изучении распределения информации в тексте.

2. Прирост и накопление информации в тексте. Выше уже говорилось (см. § 2, п. 1), что распределение информации в тексте характеризуется функцией $I_n = f(n)$. Запишем ее в виде

$$\Delta I^* = f(n),$$

где ΔI^* — прирост информации на лингвистическую единицу (букву, слог, морфему, слово и т. п.) в n-м участке текста. Сопоставляя эту зависимость с зависимостью изменения темпа развития терминологии от времени, легко заметить, что величина I_n также указывает на изменение величины информации, извлекаемой

из разных участков текста, по мере удаления соответствующего участка от начала высказывания. Правда, в отличие от возрастающего темпа развития терминологии для динамики информационного рисунка речи более характерно последовательное убывание того количества информации, которое несет каждая лингвистическая единица. Это различие, разумеется, не отражается на математической стороне наших рассуждений.

Чтобы исследовать изменения в количестве извлекаемой информации, мы можем разбить текст на равные отрезки, соответствующие, скажем, средней длине абзаца, и делать информационные замеры (например, методом угадывания; ср. [23]) в конце каждого сегмента. Но тогда от нас ускользнет динамика изменения информации внутри отрезка. Чтобы сделать наши измерения более точными, будем последовательно уменьшать интервалы измерения*, как это мы делали, рассматривая прирост терминологии**; в результате получим выражение

$$\Delta I^* = [f(n) \pm \alpha]\,\Delta n = f(n)\,\Delta n \pm \alpha \Delta n,$$

аналогичное выражению (4.21).

Рассуждая так же, как и в п. 1, приходим к выводу, что $f(n)$ есть производная некоторой функции $F(n)$, которая указывает общее количество информации I_n^*, извлеченной из текста к моменту n. Таким образом, задача снова сводится к нахождению первообразной функции $I_n^* = F(n)$ по заданной производной $I_n = f(n)$.

В будущем мы будем довольно часто встречаться с лингвистическими и лингво-статистическими зависимостями, которые, как и только что рассмотренные зависимости, могут иметь интегральную и дифференциальную формы выражения. Если для математики безразлично, определяется ли интегральная первообразная функция $F(x)$ по производной $F'(x) = f(x)$ [по дифференциалу $f(x)\,dx$] или наоборот, то для лингвистики обычно одна из этих функций является исходной. Так, в п. 1 мы исходили из статистики прироста новых терминов за единицу времени, чтобы затем, определив интегральную функцию накопления терминов, указать нужный объем соответствующего отраслевого словаря или машинного словника.

* Этот прием издавна используется в неявном виде и лингвистами. Так, например, чтобы преодолеть лингво-географическую непрерывность, диалектологи делят язык на наречия, наречия подразделяют на диалекты, диалекты дробят на поддиалекты, поддиалекты — на говоры и т. д. Сходным образом звуки (звукотипы) делятся на произносительные варианты, в которых выделяются оттенки, и т. п.

** Если речь идет о буквенном тексте или о его фонематической транскрипции, то осуществить операцию бесконечного уменьшения интервала нельзя из-за линейной (синтагматической) неделимости буквы и фонемы. Однако если речь идет об инструментальной (спектрографической, осциллографической, кимографической и т. п.) записи устной речи, то эта запись, имея непрерывный характер, допускает бесконечное уменьшение интервала (величина I_n будет в этом случае непрерывной). Разумеется, для осуществления описанного эксперимента нужно располагать некоторой процедурой измерения информации относительно непрерывного речевого процесса.

Напротив, при информационном измерении устной речи целесообразно исходить из величины накопленной информации, т. е. из интегральной функции, дифференцируя которую, можно определить темп прироста информации в тексте.

§ 4. Основные понятия интегрирования и применение их к лингвистическим задачам

1. Неопределенный интеграл. Мы только что выяснили, что функцию $f(x)$ можно рассматривать в качестве производной некоторой первообразной функции $F(x)$. При этом возникает вопрос: сколько первообразных может иметь функция $f(x)$?

Так как производная постоянной величины C равна нулю, то функции, отличающиеся друг от друга постоянными слагаемыми, имеют одну и ту же производную или один и тот же дифференциал.

Так, например, функции

$$F(x) = x^3,\ F_1(x) = x^3 + 10,\ F_2(x) = x^3 - 5,\ F_n(x) = x^3 + a$$

имеют одну и ту же производную, равную $f(x) = 3x^2$.

Обозначим постоянное слагаемое в вышеприведенных равенствах буквой C ($C = \text{const}$). Тогда получаем равенство

$$[F(x) + C]'_x = F'(x) = f(x),$$

из которого следует, что функция $f(x)$ имеет бесчисленное количество первообразных функций, отличающихся друг от друга на постоянное слагаемое C. Это множество всех первообразных функций $F(x) + C$ называется *неопределенным интегралом* и обозначается в виде

$$\int f(x)\,dx = F(x) + C. \qquad (4.23)$$

Символ $\int$ называется *знаком интеграла*, $f(x)\,dx$ — *подынтегральным выражением*, $f(x)$ — *подынтегральной функцией*, а x — *переменной интегрирования*. Левая часть равенств (4.23) читается: «неопределенный интеграл эф от икс де икс» и является общим выражением первообразной функции, в то время как правая часть этого равенства представляет собой уже найденный интеграл, где $F(x)$ — одна из первообразных функций по отношению к $f(x)$, а C — произвольная постоянная.

Так как неопределенный интеграл содержит произвольную постоянную, то он может определяться с точностью до этого произвольного слагаемого.

Пусть темп количественного роста терминологии в некоторой отрасли знаний определяется функцией $f(x)$, где x — время существования этой отрасли науки. Тогда общее количество терминов, накопленное к моменту x, выражается неопределенным интегралом:

$$y = \int f(x)\,dx = F(x) + C.$$

Величину C можно интерпретировать как некоторое количество терминологических слов и выражений, использованных данной отраслью знаний при ее возникновении из других подъязыков (например, подъязык авиации использовал при своем формировании в качестве исходного материала некоторое количество морских терминов).

При нахождении интеграла (4.23) мы получаем бесконечное множество ответов, отличающихся друг от друга на постоянное слагаемое. Неоднозначность решения можно проиллюстрировать геометрически. Для этого построим кривую, представляющую график одной первообразной функции $y = F(x)$ (при $C = 0$). Тогда остальные кривые накопления терминологии получаются в результате смещения по оси ординат этой кривой на произвольную постоянную величину C, характеризующую количество исходных терминов (рис. 34).

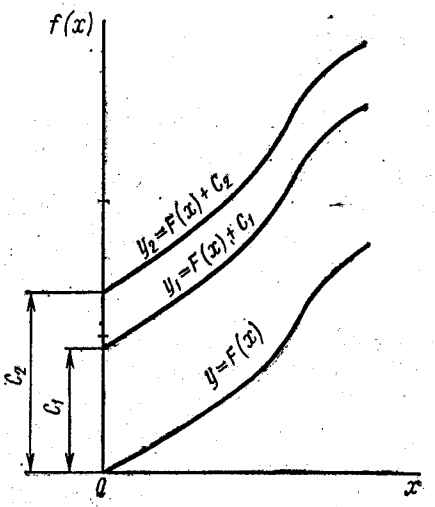


Рис. 34

Неопределенный интеграл обладает следующими свойствами.

1. *Производная от неопределенного интеграла равна подынтегральной функции*:

$$\left(\int f(x)\,dx\right)'_x = f(x).$$

2. *Дифференциал неопределенного интеграла равен подынтегральному выражению*:

$$d\int f(x)\,dx = f(x)\,dx,$$

здесь символы d и $\int$ взаимно уничтожают друг друга.

3. *Неопределенный интеграл от дифференциала функции равен этой функции плюс произвольная постоянная*:

$$\int dF(x) = F(x) + C.$$

4. *Постоянный множитель подынтегрального выражения можно вынести за знак интеграла*:

$$\int af(x)\,dx = a\int f(x)\,dx.$$

5. *Неопределенный интеграл от алгебраической суммы двух или нескольких функций равен сумме их интегралов*:

$$\int [f_1(x) + f_2(x) - f_3(x)]\,dx = \int f_1(x)\,dx + \int f_2(x)\,dx - \int f_3(x)\,dx.$$

2. Таблица простейших интегралов. Из определения неопределенного интеграла (4.23) следует, что под знаком интеграла стоит производная первообразной функции, т. е.

$$\int f(x)\,dx = \int F'(x)\,dx = F(x) + C.$$

Исходя из этого равенства, с помощью известных формул дифференцирования можно найти интегралы некоторых элементарных функций, так называемые *табличные интегралы.*

Ниже приводится таблица простейших интегралов, наиболее часто используемых в квантитативной лингвистике.

Таблица простейших интегралов

$$\int dx = x + C, \qquad \text{(I)}$$

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} + C^*, \qquad \text{(II)}$$

$$\int \frac{dx}{x} = \int x^{-1}\,dx = \ln|x| + C, \qquad \text{(III)}$$

$$\int a^x\,dx = \frac{a^x}{\ln a} + C, \qquad \text{(IV)}$$

$$\int e^x\,dx = e^x + C, \qquad \text{(V)}$$

$$\int \cos x\,dx = \sin x + C, \qquad \text{(VI)}$$

$$\int \sin x\,dx = -\cos x + C, \qquad \text{(VII)}$$

$$\int \frac{dx}{\cos^2 x} = \operatorname{tg} x + C, \qquad \text{(VIII)}$$

$$\int \frac{dx}{\sin^2 x} = -\operatorname{ctg} x + C, \qquad \text{(IX)}$$

$$\int \frac{dx}{1+x^2} = \operatorname{arctg} x + C. \qquad \text{(X)}$$

Справедливость формул интегрирования легко проверяется с помощью обратного действия — дифференцирования. При этом оказывается, что дифференциал от правой части формулы равен подынтегральному выражению.

3. Приемы интегрирования и их применение в лингвистических задачах. Нахождение интегралов, опирающееся на использование приведенных выше формул и свойств неопределенного интеграла, называется *непосредственным интегрированием.* Этот способ включает следующие случаи:

1) интеграл берется по одной из вышеприведенных формул;

* Формула II справедлива при условии, что $n \neq -1$, в противном случае знаменатель дроби в этом выражении обращается в нуль и формула теряет смысл.

2) путем использования свойств 4 или 5 неопределенный интеграл приводится к одному или нескольким табличным интегралам;

3) в результате тождественного преобразования подынтегральной функции и использования свойств неопределенного интеграла данный интеграл приводится к одному из табличных интегралов.

Чаще всего здесь применяются такие преобразования:

а) введение под знак дифференциала постоянного слагаемого

$$dx = d(x + C),$$

б) введение под знак дифференциала постоянного множителя:

$$dx = \frac{1}{m}\,d(mx),$$

в) введение под знак дифференциала постоянного множителя и слагаемого:

$$dx = \frac{1}{m}d(mx + C).$$

В лингвистике интегрирование используется обычно при решении диахронических и информационных задач.

Рассмотрим следующий диахронический пример. Пусть скорость прироста новых терминов и терминологических словосочетаний в определенной области знаний для германских, романских и славянских языков характеризуется равенством

$$f(x) = 3x^2.$$

Количество новых терминов в момент x можно определить, проинтегрировав правую часть этого равенства. Используя свойство 4 и интеграл II таблицы, находим

$$\int 3x^2\,dx = 3\int x^2\,dx = 3\,\frac{x^{2+1}}{2+1} + a = x^3 + a;$$

здесь постоянная интегрирования C заменена на a (ср. с примером в гл. 3, § 2, п. 1).

Рассмотрим теперь информационный пример. Как известно, скорость изменения величины информации, извлекаемой из разных участков текста, определяется равенством

$$I_n = (I_0 - I_\infty)\,e^{-sn} + I_\infty.$$

Определим объем информации, извлеченной из текста к моменту n (n — непрерывная величина). Для этого проинтегрируем правую часть последнего равенства:

$$\int [(I_0 - I_\infty)e^{-sn} + I_\infty]\,dn = \int I_0 e^{-sn}\,dn - \int I_\infty e^{-sn}\,dn + \int I_\infty\,dn =$$

$$= I_0 \int \frac{e^{-sn}\,d(-sn)}{-s} - I_\infty \int \frac{e^{-sn}\,d(-sn)}{-s} + I_\infty \int dn =$$

$$= -\frac{I_0 e^{-sn}}{s} + \frac{I_\infty e^{-sn}}{s} + I_\infty n + C = I_\infty n - \frac{e^{-sn}}{s}(I_0 - I_\infty) + C.$$

Если заданный интеграл нельзя или трудно привести к табличному с помощью непосредственного интегрирования, то для отыскания этого интеграла применяются специальные приемы.

Рассмотрим два из них — интегрирование способом подстановки (или замены переменной) и интегрирование по частям.

Начнем с интегрирования *способом подстановки*. Сущность этого приема состоит в том, что в интеграле вида $\int f(x)\,dx$ переменная интегрирования заменяется другой переменной u, являющейся также функцией от x. При этом данный интеграл преобразуется в новый интеграл $\int \varphi(u)\,du$, который можно вычислить с помощью одного из табличных интегралов. После того как новый интеграл взят, следует вернуться к исходной переменной x с помощью равенства $u = f(x)$.

Для иллюстрации этого метода рассмотрим интеграл

$$\int (a + x^3)^5 x^2 dx,$$

который методом непосредственного интегрирования найти нельзя. Обозначим выражение, стоящее в скобках, через u, т. е. $(a + x^3) = u^5$; следовательно, $d(a + x^3) = du$, откуда

$$3x^2\,dx = du, \text{ или } x^2\,dx = \frac{du}{3}.$$

Разбив наш интеграл на два множителя, из которых один $(a+x^3)^5 = u^5$, а другой $x^2dx = \frac{du}{3}$, мы приходим к новому интегралу:

$$\int u^5 \frac{du}{3} = \frac{1}{3}\int u^5\,du = \frac{1}{3}\,\frac{u^{5+1}}{5+1} + C = \frac{u^6}{18} + C.$$

После того как новый интеграл найден, следует вернуться к первоначальной переменной. Для этого в полученный результат вместо u подставим его значение; окончательно имеем

$$\int (a+x^3)^5\,x^2\,dx = \frac{u^6}{18} + C = \frac{(a+x^3)^6}{18} + C.$$

Если интеграл нельзя найти ни путем непосредственного интегрирования, ни с помощью замены переменной, то применяется метод *интегрирования по частям*. Пусть u и v — функции от x, т. е. $u = \varphi(x)$, а $v = \Phi(x)$. Согласно общим правилам дифференцирования (см. гл. 3, § 2, п. 1) дифференциал произведения этих функций равен

$$d(uv) = u\,dv + v\,du.$$

Интегрируя обе части этого соотношения, получаем равенство

$$\int d(uv) = \int u\,dv + \int v\,du,$$

которое, в силу свойств 3 и 5 интеграла принимает вид

$$uv + C = \int u\,dv + \int v\,du.$$

Отсюда мы приходим к формуле интегрирования по частям:

$$\int u\,dv = uv - \int v\,du + C. \qquad (4.24)$$

Левая часть этой формулы представляет собой исходный интеграл, подынтегральное выражение которого должно быть представлено в виде двух сомножителей. Подбор сомножителей u и dv следует производить таким образом, чтобы дифференцирование функции v и вычисление интеграла $\int v\,du$ представляло собой более простую задачу, чем непосредственное вычисление интеграла $\int u\,dv$. Упрощение рассматриваемого интеграла может быть достигнуто за счет дифференцирования множителя u. Поэтому ту часть подынтегрального выражения, которая упрощается при дифференцировании, следует принять за u, а все остальные сомножители подынтегрального выражения, включая dx, — за dv.

Для иллюстрации этого приема найдем интеграл $\int xe^x dx$. Полагая

$$u = x, \qquad e^x\,dx = dv,$$

имеем

$$du = dx, \qquad \int dv = \int e^x dx, \text{ т. е. } v = e^x.$$

Отсюда по формуле (4.24) получаем

$$\int xe^x\,dx = xe^x - \int e^x\,dx = xe^x - e^x + C.$$

4. Определенный интеграл и его свойства. Выше было показано, что неопределенный интеграл представляет собой бесчисленное множество первообразных функций, отличающихся друг от друга произвольным постоянным слагаемым:

$$\int f(x)\,dx = F(x) + C$$

[см. равенство (4.23)]. Поэтому неопределенный интеграл не может быть выражен определенным числом: он всегда вычисляется с точностью до произвольного слагаемого C.

Однако при решении конкретных задач бывает необходимо точно фиксировать величину C и таким образом определить начало того интервала, в котором находится переменная x и в котором осуществляется интегрирование. Предположим, что нам известен такой хронологический момент $x = a$, когда число терминов данной области знаний (как исходных, так и вновь образованных) равно нулю (этот момент существовал, очевидно, до возникновения интересующей нас отрасли знаний). Это положение можно выразить так:

$$F(a) + C = 0.$$

Следовательно, фиксированное значение C составляет

$$C = -F(a).$$

Подставив найденное значение величины C в правую часть равенства (4.23), получаем первообразную функцию в виде $F(x) - F(a)$. Тогда равенство (4.23) можно записать таким образом:

$$\int_a^x f(x)\,dx = F(x) - F(a). \qquad (4.25)$$

Здесь a — *нижний предел* интеграла (п о с т о я н н ы й), b — его *верхний предел* (п е р е м е н н ы й), а сам интеграл (4.25) называется *определенным интегралом с переменным верхним пределом*. В том случае, когда верхний предел интеграла есть постоянное число b, выражение (4.25) принимает вид

$$\int_a^b f(x)\,dx = F(b) - F(a) \qquad (4.26)$$

и называется *формулой Ньютона—Лейбница* (или *определенным интегралом с постоянными пределами*). Это читается так: «определенный интеграл от a до b эф от икс де икс».

Если неопределенный интеграл является функцией, вычисляемой с точностью до произвольного слагаемого C, то определенный интеграл есть число, указывающее на приращение первообразной функции $F(x) + C$ (т. е. неопределенного интеграла) при изменении аргумента x в интервале от a до b. Эта связь определенного и неопределенного интегралов выражена в самой формуле Ньютона—Лейбница.

Для нас в дальнейшем будут важны следующие четыре свойства определенного интеграла.

1. *При перемене местами пределов интегрирования определенный интеграл меняет свой знак на противоположный*:

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx.$$

2. *Постоянный множитель можно выносить за знак определенного интеграла*:

$$\int_a^b cf(x)\,dx = c\int_a^b f(x)\,dx.$$

3. *Определенный интеграл от алгебраической суммы двух или нескольких функций равен алгебраической сумме их интегралов*:

$$\int_a^b [f_1(x) + f_2(x) - f_3(x)]\,dx = \int_a^b f_1(x)\,dx + \int_a^b f_2(x)\,dx - \int_a^b f_3(x)\,dx.$$

4. *Промежуток интегрирования (a, b) можно разбить на несколько частичных промежутков; в этом случае интеграл, вычисленный по целому промежутку, равен сумме интегралов, вычисленных по частичным промежуткам*, т. е.

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx,$$

где $a < c < b$.

5. Вычисление определенного интеграла и численная оценка накопления новых терминов. Основным способом вычисления определенных интегралов является применение формулы (4.26). При этом здесь используются те приемы, которые применялись для вычисления неопределенных интегралов — в частности, непосредственное интегрирование и интегрирование по частям. Следует иметь в виду также, что равенство (4.26) применимо только тогда, когда промежуток интегрирования конечен, а подынтегральная функция в этом промежутке непрерывна.

Для непосредственного вычисления определенного интеграла $\int_a^b f(x)\,dx$ по формуле (4.26) нужно найти какую-либо первообразную $F(x)$ подынтегральной функции $f(x)$ и взять разность значений этой первообразной, вычисленных для значений x, равных верхней и нижним границам интегрирования.

Разность $F(b) - F(a)$ символически обозначают $F(x)\Big|_a^b$. Используя это обозначение, запишем формулу Ньютона — Лейбница в таком виде:

$$\int_a^b f(x)\,dx = F(x)\Big|_a^b = F(b) - F(a). \qquad (4.27)$$

Выше (см. п. 3) мы рассматривали вопрос об определении в некоторый момент времени x количества новых терминов, скорость прироста которых характеризуется равенством $f(x) = 3x^2$. Оценим с помощью формулы Ньютона—Лейбница число новых терминов во втором десятилетии существования интересующей нас отрасли знания.

Для этого нужно найти определенный интеграл $\int_{10}^{20} 3x^2 dx$. Так как одной из первообразных для $f(x) = 3x^2$ является функция $F(x) = x^3$, то на основании формулы (4.27) получаем

$$\int_{10}^{20} 3x^2\,dx = x^3\Big|_{10}^{20} = 20^3 - 10^3 = 8\,000 - 1\,000 = 7\,000.$$

Таким образом, за второе десятилетие существования некоторой области знаний в европейских языках может появиться около 7 тыс. новых терминов.

6. Несобственный интеграл. До сих пор мы имели дело с определенным интегралом от непрерывной функции, причем пределы интегрирования представляли собой конечные величины. Однако

могут встретиться такие определенные интегралы, у которых один или оба предела интегрирования бесконечны, или же такие, у которых подынтегральная функция имеет точки разрыва в промежутке интегрирования. Эти интегралы называются *несобственными*.

В дальнейшем нам придется иметь дело только с несобственными интегралами, имеющими бесконечные пределы интегрирования. Поэтому мы ограничимся рассмотрением именно этих интегралов, выделяя здесь три случая:

а) областью задания подынтегральной функции служит интервал $[a, +\infty)$, при этом имеет место равенство

$$\int_a^{+\infty} f(x)\,dx = \lim_{b\to+\infty} \int_a^b f(x)\,dx \qquad (4.28)$$

(рис. 35, *а*);

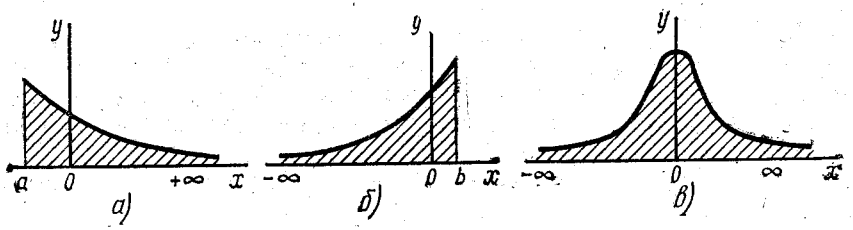


Рис. 35

б) областью задания подынтегральной функции служит интервал $(-\infty, b]$, тогда

$$\int_{-\infty}^b f(x)\,dx = \lim_{a\to-\infty} \int_a^b f(x)\,dx \qquad (4.29)$$

(рис. 35, *б*);

в) областью задания подынтегральной функции является вся числовая ось, тогда несобственный интеграл имеет вид

$$\int_{-\infty}^{+\infty} f(x)\,dx = \lim_{\substack{a\to-\infty \\ b\to+\infty}} \int_a^b f(x)\,dx \qquad (4.30)$$

(рис. 35, *в*).

В том случае, если существует предел соответствующего определенного интеграла, несобственный интеграл называется *сходящимся*, если же этот предел не существует, — *расходящимся*.

Можно показать [28], что основные свойства определенных интегралов обобщаются и на несобственные интегралы.

7. Неберущиеся интегралы. Интеграл вероятностей. Не всякий неопределенный интеграл может быть выражен через элементарные функции описанными выше способами. Среди этих неберущихся интегралов особый интерес для нас представляет *интеграл вероятностей*, который в зависимости от указанных пределов принимает вид либо определенного интеграла:

$$\int_{x_1}^{x_2} e^{-z^2/2}\,dz,$$

либо несобственных интегралов:

$$\int_{-\infty}^{+\infty} e^{-z^2/2}\,dz, \quad \int_{-\infty}^{0} e^{-z^2/2}\,dz$$

(вместо переменной x здесь и в гл. 6 используется переменная z).

Проведенное с помощью понятий математического анализа моделирование лингвистических процессов показало, что такая конфронтация языка и математики служит эффективным средством выявления скрытых от прямого лингвистического наблюдения свойств языка и речи.

Так, применение производной позволяет сформулировать и ввести в лингвистический обиход новое понятие скорости изменения в языке и речи, а также осуществлять количественные оценки этих изменений. С помощью тригонометрических функций и понятия предела моделируется циклический и ступенчатый характер лингвистических процессов (ср. понятие диахронического скачка). Ступенчатый характер диахронических и информационно-текстовых процессов, очевидно, имеет ту же природу, что и большинство кибернетических процессов. Для их описания в будущем найдут применение разделы теории обобщенных функций (например функция-ступенька). Одновременно обнаруживается необходимость описывать процессы приращения и накопления новых лингвистических элементов (диахрония) и синтактико-смысловой информации (текст), а эти процессы моделируются с помощью аппарата теории рядов и интегрирования.

Следует обратить внимание на тот факт, что как диахронические, так и текстовые информационные процессы часто аппроксимируются одними и теми же функциональными зависимостями.

Эта общность диахронических и текстовых моделей могла бы служить подтверждением гипотезы Г. Хердана [51, с. 173] о единстве лингвистического онтогенеза (информационной структуры речи) и филогенеза (формирования современного состояния языка).

ЧАСТЬ ВТОРАЯ

ВЕРОЯТНОСТНО-ИНФОРМАЦИОННЫЕ ОЦЕНКИ НОРМЫ ЯЗЫКА И СТАТИСТИЧЕСКОЕ ПОСТРОЕНИЕ ТЕКСТА

ГЛАВА 5

КОМБИНАТОРИКА ЛИНГВИСТИЧЕСКИХ ЕДИНИЦ. ВЕРОЯТНОСТЬ И ИНФОРМАЦИЯ ЛИНГВИСТИЧЕСКИХ СОБЫТИЙ

§ 1. Комбинаторные схемы

1. Измерение комбинаторики внутри лингвистических множеств. Языковеду постоянно приходится решать задачи, в которых рассматриваются комбинации и расположения элементов, принадлежащих определенному лингвистическому множеству. Так, например, синтаксисту важно знать, сколько позиционных вариантов может давать в устно-разговорной речи предложение *сегодня идет дождь*. Фонетисту, специалисту в области кодирования текста, а также работнику Госавтоинспекции, занимающемуся распределением буквенных серий автомобильных знаков на территории страны, нужно знать, сколько двух- и трехбуквенных комбинаций может дать русский алфавит. Иногда при этом нужно выяснить, какая часть этих комбинаций образует слова и их формы, использующиеся в современном русском языке. Задачи, в которых требуется ответить на вопрос «сколько?» или «сколькими способами?», называются комбинаторными, а раздел математики, занимающийся решением подобных задач, именуется комбинаторикой. Простейшие задачи комбинаторики можно решать перебором всех возможных вариантов. Так, например, путем перебора нетрудно установить, что предложение *сегодня идет дождь* имеет в русской разговорной речи 6 вариантов:

сегодня идет дождь; *сегодня дождь идет*; *дождь сегодня идет*; *дождь идет сегодня*; *идет сегодня дождь*; *идет дождь сегодня*.

Однако число комбинаций быстро растет с увеличением числа составляющих их элементов. Так, например, четыре слова (*увы*, *сегодня*, *дождь*, *идет*) дают 24, пять слов — 120, шесть — 720 позиционных вариантов и т. д. Не все из этих вариантов допустимы с точки зрения норм современного литературного языка. Определить допустимые варианты путем простого перебора оказывается невозможным.

Поэтому, сталкиваясь с такими комбинаторными задачами, прибегают к типовым схемам решения, учитывающим лингвистические или какие-либо другие ограничения.

2. Размещения. Предположим, что имеется алфавит, включающий n элементов. Из этих элементов составляются m-членные комбинации (соединения), причем каждый из n элементов может входить в соединение не более одного раза.

Такой тип комбинаций называется *размещением*. Число размещений из n элементов по m определяется по формуле

$$A_n^m = n(n-1)\dots(n-m+1) = \frac{n!}{(n-m)!}. \qquad (5.1)$$

Например, из 32 букв русского алфавита можно составить

$$A_{32}^2 = \frac{32!}{(30-2)!} = 32 \cdot 31 = 992$$

двухбуквенные комбинации, не содержащие повторений букв.

По данным четырехтомного «Словаря русского языка» (М., 1957—1961), из этих сочетаний только 114 выступает в качестве самостоятельных слов (имена собственные, сокращения, архаизмы и диалектные слова при этом не учитываются).

3. Размещения с повторениями. Снова возьмем алфавит из n элементов и будем составлять m-членные соединения, допуская повторения каждого элемента от 0 до m раз. Тогда общее число соединений, называемых *размещениями с повторениями*, находится по формуле

$$\tilde{A}_n^m = n^m. \qquad (5.2)$$

Так, например, из 30 букв русского алфавита (исключая ь и ъ) можно составить $30^2 = 900$ двухбуквенных серий для денежных знаков и $30^3 = 27\,000$ трехбуквенных серий для автомобильных номеров.

4. Перестановки. Пусть размещения из n разных элементов взяты по n элементов, т. е. каждое размещение содержит все n элементов алфавита и отличается от других лишь порядком этих элементов. Такие размещения называются *перестановками*. Тогда из формулы (5.1) можно получить формулу для нахождения числа перестановок, заменив m на n и учитывая, что $0! = 1$. Действительно,

$$A_n^n = P_n = \frac{n!}{(n-n)!} = n!. \qquad (5.3)$$

Определим, например, сколько трехсловных предложений можно построить из трех слов: *сегодня*, *идет*, *дождь*. Число предложений равно здесь числу перестановок из трех элементов: $P_3 = 2 \cdot 3 = 6$. К этому же результату мы пришли в п. 1, используя метод простого перебора.

5. Перестановки с повторениями. В тех случаях, когда среди образующих перестановки элементов имеются одинаковые, полу-

чаются соединения, называемые *перестановками с повторениями*. Число этих перестановок вычисляется по формуле

$$P^{n}_{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1!\, n_2! \ldots n_k!}, \qquad (5.4)$$

где n — общее количество элементов, входящих в перестановку, а n_1, n_2, ..., n_k — количество одинаковых элементов в первой, второй, ..., k-й группах.

Определим, например, число перестановок с повторениями, которое можно получить из букв, составляющих словоформу *математика*. Всего в перестановках участвует десять букв, т. е. $n = 10$; буква *м* повторяется два раза, поэтому если бы все остальные буквы были различными, то искомое число перестановок, было бы равно $P^{10}_2 = 10!/2!$. На самом деле, кроме двух одинаковых *м* в нашем слове имеются три *а* и два *т*. Поэтому общее число перестановок, полученных из букв, входящих в словоформу *математика*, равно

$$P^{10}_{2,2,3} = \frac{10!}{2!\,3!\,2!} = \frac{1\cdot2\cdot3\cdot4\cdot5\cdot6\cdot7\cdot8\cdot9\cdot10}{1\cdot2\cdot1\cdot2\cdot3\cdot1\cdot2} = 5\cdot6\cdot7\cdot8\cdot9\cdot10 = 151\,200.$$

Кстати говоря, среди более чем ста пятидесяти тысяч десятибуквенных комбинаций, составленных из двух *м*, трех *а*, двух *т* и *е*, *к*, *и*, только одна — *математика* — является «отмеченной» в системе русского языка. Остальные оказываются лишенными смысла, избыточными с точки зрения современного русского языка последовательностями букв.

6. Сочетания. В размещениях из n элементов по m соединения отличаются друг от друга либо элементами, либо их порядком, либо и элементами и их порядком. Объединим в отдельные группы такие комбинации, которые содержат m одинаковых элементов и отличаются друг от друга только порядком этих элементов. Нетрудно заметить, что в каждой группе будет ровно P_m элементов. Группы комбинаций, различающиеся только элементами, называются *сочетаниями* из n элементов по m. Их число равно

$$C^m_n = \frac{n!}{m!\,(n-m)!} = \frac{P_n}{P_m\, P_{n-m}} = C^{n-m}_n. \qquad (5.5)$$

Например, если имеются три согласных и две гласных фонемы, то, исходя из равенства (5.5), можно построить

$$C^2_5 = \frac{1\cdot2\cdot3\cdot4\cdot5}{1\cdot2\cdot1\cdot2\cdot3} = \frac{1\cdot2\cdot3\cdot4\cdot5}{1\cdot2\cdot3\cdot1\cdot2} = 10$$

пятифонемных «слов», отличающихся друг от друга только расположением гласных и согласных фонем.

7. Сочетания с повторениями. *Сочетаниями* из n элементов по m *с повторениями* называются такие соединения, которые включают m из n различающихся между собой элементов при условии, что один и тот же элемент может включаться в комбинацию

несколько раз. Два соединения считаются различными, если они отличаются хотя бы одним элементом, и одинаковыми, если они состоят из одних и тех же элементов. Число сочетаний из n элементов по m с повторениями определяется по формуле

$$\widetilde{C}^m_n = C^m_{n+m-1}. \qquad (5.6)$$

Рассмотрим в связи с этим следующий пример. В некотором языке имеются два типа фонем: гласные и согласные, причем слово может быть образовано из одних гласных, из одних согласных, а также из гласных и согласных (таким образом, согласные, так же как и гласные, являются слогообразующими). Необходимо определить, сколькими способами можно образовать трехфонемное «слово».

Поскольку здесь $n = 2$, а $m = 3$, то на основании соотношения (5.6) искомое число равно

$$\widetilde{C}^3_2 = C^3_{2+3-1} = \frac{4\cdot3\cdot2}{1\cdot2\cdot3} = 4.$$

Действительно, при построении трехфонемного слова возможны два случая: а) «слово» составлено из фонем одного типа; б) в «слово» входят и гласные, и согласные. В первом случае могут быть два способа образования «слова»: оно состоит либо из гласных, либо из согласных. Во втором случае также имеются два способа: либо «слово» образовано из одной гласной и двух согласных, либо из двух гласных и одной согласной. Итак, существуют только четыре способа образования трехфонемных слов.

В только что рассмотренном примере мы имеем дело с сочетаниями из двух элементов по три с повторениями.

§ 2. Лингвистическое событие

1. Наблюдение, испытание и событие в индуктивных исследованиях языка и речи. Основой всех индуктивных исследований в языкознании является наблюдение за поведением и признаками изучаемых лингвистических объектов. Это наблюдение может осуществляться также путем опыта, эксперимента или количественного измерения. Осуществление каждого такого наблюдения (опыта или измерения) называется *испытанием*. Совокупность условий, при которых осуществляется данное испытание, называют *комплексом условий* (σ).

Результатом лингвистического испытания является лингвистическое событие.

Проведем опыт (испытание), состоящий в угадывании буквы при следующем комплексе условий (σ_1): угадываемой букве предшествует цепочка *Δкоторо*, текст русский без ошибок и опечаток. Это испытание может дать события A_1, B_1, C_1, D_1, состоящие соответственно в появлении следующих букв: *г* (*которого*), *е* (*которое*), *й* (*которой*), *м* (*котором*, *которому*).

Каждое событие, которое здесь может произойти, а может и не произойти, называется *случайным событием* (ср. события A_1, B_1, C_1, D_1). Если результат лингвистического испытания полностью

исчерпывается каким-либо одним (и только одним) событием, то мы имеем дело с *элементарным* случайным событием. Событие, состоящее из нескольких элементарных событий, определяется как *сложное* случайное событие. Появления букв *г, е, й, м* после цепочки *Δкоторо* являются элементарными случайными событиями, появления после этой же цепочки диграмм *го, му* следует рассматривать как сложные случайные события.

2. Соотношения между лингвистическими событиями. Поскольку между алгеброй событий и теорией множеств существует тесная связь, мы будем пользоваться, рассматривая операции над событиями, теоретико-множественными аналогиями. Соотношения между событиями будут иллюстрироваться часто теми же рисунками, с помощью которых эксплицировались операции над множествами (см. рис. 4 на стр. 14).

1. Сложное событие, заключающееся в наступлении хотя бы одного из событий A и B, называется *суммой* этих событий и обозначается $A + B$ или $A \cup B$ (читается: «A или B»). Теоретико-множественным аналогом суммы событий является объединение множеств (см. рис. 4, *д*). Появление буквы *г* (событие A_1) или буквы *е* (событие B_1) после цепочки *Δкоторо* является суммой $A_1 + B_1$.

2. Сложное событие, состоящее в одновременном наступлении A и B, называется их *произведением* и обозначается AB или $A \cap B$ (читается: «A и B»). В качестве аналога сложного события можно рассматривать пересечение множеств (см. рис. 4, *ж*). Пусть, например, используется комплекс условий (σ_2), заключающийся в том, что русский алфавит считается состоящим из гласных и согласных. При этом буква *й* считается принадлежащей одновременно и к классу гласных, и к классу согласных. Будем считать появление гласной после цепочки *Δкоторо* событием Φ_1, а появление согласной — событием H_1. Тогда появление буквы *й* после цепочки *Δкоторо* следует рассматривать как произведение $\Phi_1 H_1$ (или $\Phi_1 \cap H_1$) данных лингвистических событий.

3. Событие, заключающееся в том, что событие A имеет место, а B не имеет места, называется *разностью* событий A и B и обозначается $A—B$. Разности событий соответствует разность множеств (см. рис. 4, *з*). Появление всех согласных, кроме *й*, после цепочки *Δкоторо* можно представить как разность $H_1 - \Phi_1$.

4. Если событие A, происходящее при реализации комплекса (σ), влечет за собой каждый раз событие B, то говорят, что A является *частным случаем* B, и записывают $A \subset B$ (или $B \supset A$). Этому соотношению в теории множеств соответствует включение (см. рис. 4, *в*). Так, например, появление буквы *г* после цепочки *Δкоторо* одновременно означает (влечет за собой) появление согласной, иными словами, здесь $A_1 \subset H_1$ (или $H_1 \supset A_1$).

5. Если событие A при комплексе условий (σ) влечет за собой событие B и, наоборот, при этом же комплексе условий B влечет A, то события A и B называют *равносильными* и записывают $A = B$ (см. рис. 4, *е*). Так, например, условившись считать появление четвертой буквы русского алфавита событием A_1' и сохраняя комплекс условий (σ_1), мы можем считать события A_1 и A_1' равносильными, записав при этом, что $A_1 = A_1'$ (это соответствует равенству: *г* = четвертая буква русского алфавита).

6. Если некоторое событие при данном комплексе условий должно непременно произойти, то такое событие называется *достоверным*. Событие, которое при комплексе условий (σ) произойти не может, называется *невозможным*. Поскольку все достоверные события равносильны между собой, их принято обозначать буквой U, невозможные события в силу этих же соображений обозначаются буквой V; $V = \overline{U}$. Проведем снова опыт по угадыванию буквы. Комплекс условий (σ_3) отличается от комплекса условий (σ_1) только тем, что угадываемой букве предшествует цепочка *Δкоторог*. Испытание дает здесь только одно событие, заключающееся в появлении буквы *о*. Это событие является достоверным. Появление любой другой буквы после цепочки *Δкоторог* представляет собой невозможное событие.

7. Два события называются *несовместимыми*, если появление одного из них при данном испытании исключает возможность появления другого. События, состоящие в появлении после цепочки *Δкоторо* букв *г* и *е*, являются несовместимыми.

8. Два события являются *совместимыми*, если появление одного из них при данном испытании не исключает появление другого. Так, например, события Φ_1 и H_1 (см выше) являются совместимыми.

9. События $A, B, C, .., Z$ образуют *полную систему событий*, если при осуществлении испытания при комплексе условий (σ) хотя бы одно из них должно произойти. События, состоящие в появлении после цепочки *Δкоторо* букв *г, е, й, м*, образуют полную систему событий.

10. Два несовместимых события A и $\bar{A}$ ($\bar{A}$ читается: «не A»), составляющих полную систему событий, называются *противоположными*. Угадывая букву после цепочки *Δкоторо*м, имеем два противоположных события, образующих полную систему. Первое из них состоит в появлении буквы *у* (*которому*), второе заключается в появлении пробела Δ (*котором* Δ)

Система, включающая простые события $A, B, C, ...$, а также сложные события, представляющие суммы, произведения, разности, отрицания и т. п., называется *полем событий*.

§ 3. Вероятность элементарного лингвистического события

Простое перечисление и классификация лингвистических событий, которые образуют поле событий, принадлежащее данному опыту, имеет сравнительно ограниченный познавательный интерес. Гораздо важнее оценить степень возможности того или иного события.

Мерой возможности появления лингвистического события A при осуществлении комплекса условий (σ) является вероятность $P(A)$ этого события. Большинство определений вероятности носит в боль-

шей или меньшей степени операционный характер, т. е. опирается на конкретный прием оценки вероятности того или иного события. Для языкознания интерес представляют три определения вероятности: а) определение вероятности, исходящее из субъективной количественной оценки возможности события; б) «классическое» определение вероятности; в) «статистическое» определение вероятности.

1. Субъективное определение вероятности и его использование в лингвистике. Если человек решается интуитивно оценить вероятность события A, то он опирается на совокупность знаний (тезаурус) Θ относительно тех возможностей, которые могут способствовать или не благоприятствовать осуществлению события A.

Эта вероятность может быть представлена как $P(A, \Theta)$, т. е. как вероятность события A при заключенном в мозгу данного человека тезаурусе Θ. Если два человека имеют относительно события A одинаковый тезаурус Θ, то значения вероятностей события A для этих людей будут одни и те же [14, с. 13]. Однако такая ситуация встречается редко. Чаще вероятность одного и того же события оценивается разными людьми, исходя из разных величин Θ, Θ'. Даже у одного и того же познающего субъекта со временем величина Θ изменяется и превращается в Θ', следовательно, и его оценки вероятности события A в разные периоды его жизни являются различными: $P(A, \Theta) \neq P(A, \Theta')$.

Так, например, человек, недостаточно знающий русский язык, может предполагать, что вероятности появлений букв *г*, *е*, *й*, *м* после цепочки Δ*которо* равны. Напротив, человек, хорошо знающий русский язык и ориентирующийся на художественную прозу и разговорную речь, скажет, что вероятность появления *е* или *й* после указанной цепочки выше, чем появление *г* или *м*. Наконец, информант, языковое чутье которого сформировалось на основе газетной речи, будет утверждать, что наиболее вероятной в данной ситуации является* буква *г* [6, с. 46].

Часто говорят, что оценка вероятности того или иного события имеет отношение только к состоянию познающего субъекта, и поэтому все выводы из вероятностных суждений лишаются объективного, не зависящего от познающего субъекта содержания [14, с. 18]. Вместе с тем нельзя забывать, что многие исследования в области экспериментальной психологии и языкознания строятся на основе

* На использовании субъективных вероятностей строятся многие языковедческие исследования, а различия в субъективных вероятностях становятся часто источником разного вида лингвистических дискуссий.

Так, например, А. Вайан считал, что русский именной суффикс-*яга* (ср. *бродяга*, *работяга*, *стиляга*) продуктивен [60, с. 77—79], т. е. вероятность образования с ним новых слов достаточно велика; В. В. Виноградов, наоборот, утверждает, что этот суффикс малопродуктивен, т. е. вероятность появления с ним новых слов очень мала [11, с. 75]. А. Н. Гвоздев считал, что образования с приставкой *раз-*(*разудалый*, *развеселый*) вероятнее всего встречаются в разговорном языке и просторечии [6, с. 46, прим.], а Академическая Грамматика утверждает, что эти слова вероятнее всего можно встретить в диалектах [16, с. 358].

обработки именно субъективных вероятностных оценок, получаемых исследователем от информанта [21]; [23]. С помощью субъективных вероятностей оценивается достоверность вхождения объектов в нечеткие множества (см. гл. 1, § 1, п. 1). На этой же основе делаются попытки измерения семантической информации (см. ниже § 5, п. 7).

2. Классическое определение вероятности (схема случаев) и построение частотного словаря целостного текста. Существуют испытания, для которых вероятности их исходов можно оценить непосредственно из условий самого опыта. Для этого необходимо, чтобы различные исходы испытаний обладали симметрией и в силу этого были бы равновозможными.

Для иллюстрации симметрии и равновозможности опытов (схемы случаев) рассмотрим слово *кот*, составленное из букв разрезной азбуки. Карточки с буквами тщательно перемешивают и кладут в урну. Производится испытание, состоящее в извлечении карточки с буквой. Появления любой из букв, образующих слово *кот*, в силу правила симметрии, являются равновозможными и попарно несовместимыми событиями.

Теперь обратимся к некоторому множеству S попарно несовместимых равновозможных лингвистических событий $A_1, A_2, \ldots, A_N$, которые составляют полную систему событий. Образуем систему Ω, состоящую из невозможного события V, всех событий A_i множества S, а также всех комбинаций событий A_i, входящих в это множество. Если, к примеру, множество S состоит из трех событий A_1, A_2, A_3 (ср. слово *кот*), то система Ω включает события: V, A_1, A_2, A_3, $A_1 + A_2$, $A_2 + A_3$, $A_1 + A_3$, A_1A_2, A_2A_3, A_1A_3, $A_1A_2A_3$, а также $A_1 + A_2 + A_3 = U$.

Нетрудно видеть, исходя из определения, данного в п. 2 § 2, что Ω есть поле событий. Действительно, невозможное событие V входит в Ω по определению, комбинации событий A_i входят в Ω также по определению, достоверное событие U входит сюда, поскольку $U = A_1 + A_2 + \ldots + A_N$.

Для событий системы Ω может быть дано так называемое к л а с с и ч е с к о е о п р е д е л е н и е вероятности, которое формулируется следующим образом.

Если результаты испытания можно представить в виде полной системы N равновозможных и попарно несовместимых событий и если случайное событие появляется только в F случаях, то *вероятность* события A равна

$$P(A) = F/N, \tag{5.7}$$

т. е. отношению количества случаев, благоприятствующих данному событию, к общему числу всех случаев.

В нашем примере вероятность появления согласной составляет P (согл.) $= 2/3$.

Из классического определения вероятности вытекают такие с л е д с т в и я.

1. *Вероятность достоверного события равна единице*:

$$P(U) = 1.$$

2. *Вероятность невозможного события равна нулю*:

$$P(V) = 0.$$

3. *Вероятность появления случайного события A при N испытаниях есть положительное число, заключенное между нулем и единицей*:

$$0 \leqslant P(A) \leqslant 1.$$

В некоторых лингвистических работах, использующих элементы теории вероятностей, величина вероятности выражается в процентах.

Исходя из классического определения вероятности, осуществляется вероятностная обработка частотных словарей отдельных произведений или всего творчества писателя. В этих случаях все словоупотребления, составляющие текст всех произведений или отдельного произведения, подчиняются правилу симметрии и образуют полную систему равновозможных и попарно несовместимых событий. Некоторое интересующее нас слово (или словоформа) A появляется в исследуемом тексте в виде словоупотреблений. Отсюда вероятность того, что наугад взятое слово из нашего текста окажется именно словом (словоформой) A, согласно (5.7), равна $P(A) = F/N$.

Например, текст «Капитанской дочки» А. С. Пушкина состоит из 29343 словоупотреблений. Формы слова *быть* встречаются здесь 430 раз. Отсюда следует, что вероятность появления в тексте «Капитанской дочки» форм слова *быть* такова:

$$P_1 \ (быть) = F_1/N_1 = 430/29343 \simeq 0{,}0147 = 1{,}47\%.$$

Что касается всего корпуса текстов А. С. Пушкина, который состоит из 544777 словоупотреблений, то здесь формы слова *быть* употреблены автором 8771 раз. Вероятность того, что наугад взятое слово окажется словом *быть* в любом произведении Пушкина, составит [6, с. 51—52]

$$P_2 \ (быть) = F_2/N_2 = 8771/544777 \simeq 0{,}0161 = 1{,}61\%.$$

3. Статистическое определение вероятности. Выборочное частотное описание текста. Классическое определение вероятности оказывается весьма удобным применительно к таким опытам, которые заведомо дают симметрию конечного числа равновероятных исходов. Однако при переходе от этих простых примеров к решению более сложных вероятностно-лингвистических задач это определение наталкивается на непреодолимые трудности.

Во-первых, число возможных результатов может и не быть конечным. Так, например, определяя вероятности появления в языке слов, словоформ или сочетаний, мы должны согласиться с тем, что практически число этих лингвистических единиц стремится к бесконечности.

Во-вторых, утверждать о равновероятности исходов лингвистического опыта обычно бывает весьма затруднительно.

К опытам, которые не могут быть исследованы на основе системы случаев, применяется так называемое статистическое определение вероятности.

Прежде чем давать статистическое определение вероятности, введем некоторые определения и рассмотрим конкретный лингвистический пример.

Пусть произведена серия из N испытаний, в каждом из которых могло появиться или не появиться событие A. Тогда *абсолютной частотой* (или *частотой*) F называется число появлений события A, а *относительной частотой* (или *частостью*) $f(A)$ — отношение абсолютной частоты к общему числу испытаний:

$$f(A) = F/N. \tag{5.8}$$

При небольшом числе опытов частоты события носят непостоянный и случайный характер и могут изменяться от одной группы событий к другой. Например, в одном взятом наугад тексте из произведений Пушкина длиной в 100 слов формы глагола *быть* не появились ни разу, зато в другом отрывке той же длины этот глагол появился три раза и его относительная частота возросла до 0,03. Однако при последовательном увеличении объема выборки относительная частота глагола *быть* приобретает определенную устойчивость, приближаясь к величине 0,01 (см. табл. 5.1). Аналогичным образом получены относительные частоты (статистические вероятности) русских букв, показанные в табл 5.2.

Таблица 5.1

Относительная частота глагола *быть* в русской художественной прозе (Пушкин, Тургенев, Бунин)

Объемы выборок	10	100	1000	2000	3000	4000	5000
F	0	3	15	17	31	33	47
f	0,000	0,030	0,015	0,008	0,010	0,008	0,009

Продолжение табл. 5.1

Объемы выборок	6000	7000	8000	9000	10000	15000	400000
F	57	71	74	88	95	153	4186
f	0,010	0,010	0,009	0,010	0,010	0,010	0,011

Таблица 5.2

Распределение вероятностей букв в русских литературных текстах

Буква	P	Буква	P	Буква	P
Пробел (Δ)	0,174	*к*	0,028	*ч*	0,012
о	0,090	*м*	0,026	*й*	0,010
е, ё	0,072	*д*	0,025	*х*	0,009
а	0,062	*п*	0,023	*ж*	0,007
и	0,062	*у*	0,021	*ш*	0,006
н	0,053	*я*	0,018	*ю*	0,006
т	0,053	*в*	0,016	*ц*	0,004
с	0,045	*ы*	0,016	*щ*	0,003
р	0,040	*б*	0,014	*э*	0,003
в	0,038	*ь, ъ*	0,014	*ф*	0,002
л	0,035	*г*	0,013		

Опыт многих наук, да и вся практическая деятельность человека показывают, что результаты отдельных статистических испытаний могут давать заметные флуктуации. Однако при большом числе испытаний N статистические флуктуации начинают сглаживаться, а относительная частота f обнаруживает все большую устойчивость. Иными словами, в случайных явлениях имеется некоторое объективно существующее свойство, которое имеет тенденцию оставаться постоянным и проявляется все яснее при увеличении объема исследуемого материала. Указанное свойство измеряется некоторой постоянной величиной, которая является количественной объективной числовой характеристикой изучаемого явления. Эта постоянная величина и называется *вероятностью* случайного события A [будем по-прежнему обозначать ее символом $P(A)$]. Экспериментальными значениями вероятности являются относительные частоты интересующего нас события $f(A)$ в определенных сериях наблюдений. Определенная таким образом вероятность случайного события носит название статистической вероятности.

Следует обратить внимание читателя на то, что точное численное значение статистической вероятности остается, вообще говоря, неизвестным. За численное значение вероятности обычно принимается при большом количестве испытаний либо сама частость события A, либо некоторое число, близкое к этой частости, например некоторое среднее арифметическое относительных частот, полученных из нескольких достаточно больших серий испытаний*.

Оставляя в стороне методологические дискуссии, связанные со статистическим определением вероятности [14, с. 17 и сл.], необходимо подчеркнуть, что этот подход имеет принципиальное значение для

* Разумеется, это не значит, что вероятность события вообще не может быть точно определена. Если мы имеем дело со схемой случаев, то вероятность вычисляется по формуле (5.7). Кстати, если нас интересует вероятность появления глагола *быть* только в произведениях Пушкина, то, используя классическое определение вероятности, нетрудно показать, что она равна 0,0161.

прикладных исследований, в том числе и лингвистических, например при составлении частотных словарей. Не имея обычно возможности обследовать всю генеральную совокупность возможных исходов (например, всю совокупность словоупотреблений, составляющих все когда-либо написанные русские тексты), мы вынуждены производить серию наблюдений, охватывающих некоторую частную совокупность (например, определенную выборку из русских текстов). В результате таких исследований мы получаем относительные частоты для случайных событий (в нашем случае —словоформ или слов). По этим относительным частотам необходимо оценить численные значения вероятностей, которые, как уже указывалось, являются числовой характеристикой изучаемых явлений. Эта оценка сводится к выяснению того, насколько далеко отклоняются экспериментальные частоты от вероятности. Решение такой задачи является по существу узловым вопросом всех статистических исследований.

4. Аксиоматическое построение теории вероятностей. Все только что рассмотренные определения вероятности имеют существенные недостатки и ограничения.

Интуитивная оценка вероятности зависит от тезауруса Θ познающего субъекта, который обычно не поддается измерению. Схема случаев применима лишь к таким опытам, которые заведомо дают симметрию конечного числа равновероятных исходов. При статистическом подходе понятие вероятности вообще остается в тени.

Широкое проникновение вероятностно-статистических исследований в естественные и гуманитарные исследования потребовало создания формально-логического обоснования всего аппарата теории вероятностей; это обоснование дано в аксиоматическом построении теории вероятностей, предложенном А. Н. Колмогоровым [19].

Поскольку аксиоматика Колмогорова позволяет преодолеть ряд трудностей, возникающих при использовании теории вероятностей в языкознании, мы рассмотрим ее основные идеи и положения.

В аксиоматике Колмогорова случайное событие не рассматривается как исходное первичное понятие, но образуется на основе других элементарных понятий. Чтобы пояснить это положение, рассмотрим два примера.

Пусть имеется некоторое пространство U (прямая, площадь и т. д.). В этом пространстве содержатся подобласти A, B, ..., Z. В пространстве U «наудачу» берется точка a. Попадания точки a в те или иные подобласти точек и являются случайными событиями. Одновременно каждое случайное событие выступает в качестве некоторого подмножества множества точек U (рис. 36).

Возьмем другой пример. Пусть имеется текст, написанный на некотором естественном языке. Этот текст можно рассматривать как некоторое лингвистическое пространство (множество словоупотреблений) U. Подобластями (подмножествами) A, B, ..., Z этого пространства являются группы словоупотреблений, имеющие абсолютно одинаковое написание (т. е. словоупотребления, реализующие одну и ту же словоформу), например: *а*, *абажур*, *абажура*, *абажуру* ,..., *наука*, *науки*, *науке*, ...

Из словаря берется некоторая словоформа, например *науке*, и накладывается наугад на одно из словоупотреблений текста. Словарная единица *науке* может совпасть с текстовым словоупотреблением *науке* (в этом случае мы имеем дело с попаданием словарной единицы в подмножество *науке*), а может и не совпасть с указанным текстовым словоупотреблением. Попадание или непопадание словарной единицы в то или иное подмножество является случайным событием. При этом каждое случайное событие является некоторым подмножеством нашего лингвистического множества.

Иными словами, аксиоматика Колмогорова исходит из множества U *элементарных событий* (в наших примерах — геометрических точек или словоупотреблений), которые в данной ситуации можно рассматривать как возможные события. Далее вводится система $\mathfrak{F}$ подмножеств множества U. Элементы этой системы называются случайными событиями. Построение системы $\mathfrak{F}$ должно отвечать следующим требованиям:

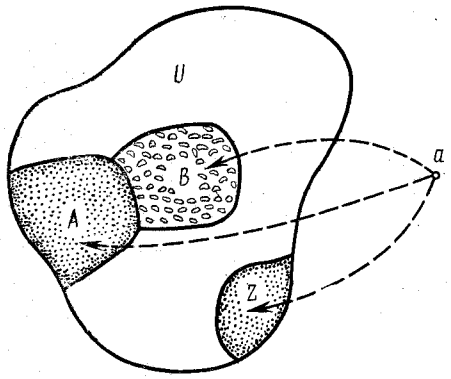


Рис. 36

1) $\mathfrak{F}$ содержит в качестве элементов множество U, а также пустое множество $\bar{U} = V$;

2) если A и B, являющиеся подмножествами множества U, входят в $\mathfrak{F}$ как его элементы, то множества $A + B$, AB, $A - B$, $\bar{A}$ и $\bar{B}$, составленные из элементов U, также будут элементами системы $\mathfrak{F}$. В этом случае $\mathfrak{F}$ называется *телом событий*.

Нетрудно заметить, что представленные здесь требования аналогичны условиям, на которых строилось поле событий Ω в классическом определении вероятности, сводящемся к схеме случаев. Однако к полю событий в аксиоматике Колмогорова предъявляется еще одно требование, позволяющее применять ее к речевым ситуациям, в которых число исходов опыта не является конечным. Это требование можно сформулировать так:

Если подмножества A_1, A_2, ..., A_n, ..., принадлежащие множеству U, суть элементы системы $\mathfrak{F}$, то их сумма $A_1 + A_2 + \ldots + A_n + \ldots$ и произведение $A_1 A_2 \ldots A_n \ldots$ также являются элементами $\mathfrak{F}$.

Изложив основную идею аксиоматики Колмогорова, перечислим теперь основные аксиомы, определяющие вероятность.

1°. *Каждому случайному событию A из поля событий $\mathfrak{F}$ можно поставить в соответствие неотрицательное число $P(A) \geqslant 0$, называемое его вероятностью.*

2°. *U является событием с вероятностью $P(U) = 1$.*

3°. (Аксиома сложения.) *Если события A и B несовместимы, то*

$$P(A + B) = P(A) + P(B). \tag{5.9}$$

Аналогично, *если события A_1, A_2, ..., A_n попарно несовместимы, то*

$$P(A_1 + A_2 + \ldots + A_n) = P(A_1) + P(A_2) + \ldots + P(A_n). \tag{5.10}$$

4°. (Аксиома непрерывности.) *Если имеется последовательность событий A_1, A_2, ..., A_n и эти события не могут осуществляться одновременно*, то

$$\lim_{n \to \infty} P(A_1 A_2 \ldots A_n) = 0.$$

5°. (Расширенная аксиома сложения.) *Аксиома сложения справедлива для бесконечного количества событий.* Иными словами, если

$$A = A_1 + A_2 + \ldots + A_n + \ldots$$

является событием, то

$$P(A) = P(A_1) + P(A_2) + \ldots + P(A_n) + \ldots . \tag{5.11}$$

Приведенные аксиомы дают ряд важных для лингвистических приложений следствий.

1. *Если несовместимые события A_1, A_2, ..., A_n образуют полную группу событий, то* согласно аксиомам 2° и 3°, *сумма вероятностей этих событий равна единице*, т. е.

$$P(A_1) + P(A_2) + \ldots + P(A_n) = 1. \tag{5.12}$$

2. *Сумма вероятностей двух противоположных событий равна единице*, т. е.

$$P(A) + P(\bar{A}) = 1. \tag{5.13}$$

3. *Из формулы* (5.13) *следует, что*

$$P(A) = 1 - P(\bar{A}),$$

или

$$P(\bar{A}) = 1 - P(A).$$

4. *Вероятность невозможного события равна нулю:*

$$P(V) = 0.$$

5. *Каково бы ни было случайное событие A, его вероятность заключена между нулем и единицей:*

$$0 \leqslant P(A) \leqslant 1. \tag{5.14}$$

6. *Если $A \subset B$, то $P(A) \leqslant P(B)$.*

7. *Если A и B — совместимые события, то в суммах $A + B = A + (B - AB)$ и $B = AB + (B - AB)$ слагаемые правых частей — несовместимые события.*

8. *Если A и B — произвольные события* (т. е. такие события, которые могут быть и совместимыми, и несовместимыми), *то имеет место неравенство*

$$P(A + B) \leqslant P(A) + P(B). \qquad (5.15)$$

По индукции следует, что *если $A_1, A_2, \ldots, A_n$ — произвольные события, то*

$$P(A_1 + A_2 + \ldots + A_n) \leqslant P(A_1) + P(A_2) + \ldots + P(A_n).$$

Приведенные аксиомы и следствия мы проиллюстрируем в дальнейшем по ходу описания лингвистических приложений теории вероятностей.

Аксиоматическое построение основ теории вероятностей характеризуется следующими особенностями:

I. Вероятностные понятия получают здесь теоретико-множественную интерпретацию. Сущность ее состоит в том, что все возможные для данного опыта элементарные события, их суммы и произведения, а также невозможные события рассматриваются как элементарные множества $\mathfrak{F}$, причем каждому элементу этого множества ставится в соответствие некоторое число (норма), являющееся его счетно-аддитивной (т. е. способной к арифметическому сложению), неотрицательной мерой. Такая интерпретация вероятностных понятий принципиально важна для лингвистического приложения теории вероятностей: она позволяет перебросить естественный мост между алгебраическим (по своей природе теоретико-множественным) языкознанием, квантитативной (по своему существу — вероятностной) лингвистикой и нечетко-множественным языкознанием [26, с. 207—269]; [65].

II. Аксиоматика Колмогорова исходит из свойств и понятий вероятности, сформулированных уже в классическом и статистическом ее определениях. Эти последние целиком включаются в аксиоматическое определение вероятности как ее частные случаи. Вместе с тем аксиоматика Колмогорова преодолевает ограниченность как классического, так и статистического определений. С одной стороны, удается избежать тех логических трудностей, которые связаны с несовместимостью понятий иррегулярности и требования о существовании предела, — понятий, постулируемых в статистическом определении вероятности. С другой стороны, преодолевается ограниченность схемы случаев, оперирующей лишь с конечным числом результатов. Введение А. Н. Колмогоровым в определение вероятности аксиомы непрерывности, а также расширенной аксиомы сложения позволяет рассматривать события, подразделяющиеся на бесконечное число частных случаев. Этот факт имеет принципиальное значение для языкознания, которое постоянно имеет дело с речевыми процессами, охватывающими бесконечное число словоформ, словосочетаний и предложений.

§ 4. Вероятности сложных лингвистических событий

1. Сложение вероятностей. Языковеда редко интересуют элементарные события, чаще всего ему приходится иметь дело со сложными лингвистическими событиями, например с суммой элементарных событий. Выбор правил, с помощью которых вычисляется вероятность сложного события, определяется тем, являются ли составляющие его элементарные события несовместимыми или совместимыми.

Согласно правилам 3^0 (аксиома сложения) и 5^0 (расширенная аксиома сложения), вероятность наступления одного из попарно независимых событий $(A_1 + A_2 + \ldots + A_n + \ldots)$ равна сумме вероятностей этих событий:

$$P(A_1 + A_2 + \ldots + A_n + \ldots) = P(A_1) + P(A_2) + \ldots + P(A_n) + \ldots .$$

Однако если два события совместимы, то их вероятность определяется как сумма вероятностей этих событий минус произведение вероятностей этих событий:

$$P(A + B) = P(A) + P(B) - P(A)\,P(B) \qquad (5.16)$$

При вычислении вероятности суммы нескольких совместимых событий обычно пользуются правилом, согласно которому вероятность появления хотя бы одного из нескольких совместимых событий $A_1, A_2, \ldots, A_n$ равна разности между единицей и вероятностью совместного наступления (умножения) всех противоположных событий. Иными словами,

$$P(A_1 + A_2 + \ldots + A_n) = 1 - P(\bar{A}_1 \bar{A}_2 \ldots \bar{A}_n) = 1 - \prod_{i=1}^{n} (1 - P(A_i)). \qquad (5.17)$$

2. Прогнозирование вероятностей лингвистических событий при повторении опытов. Рассмотренные правила широко используются при прогнозировании событий в разного рода вероятностно-лингвистических, инженерно-лингвистических и информационных задачах. Рассмотрим в этой связи следующий пример.

Для расчета памяти вероятностного автомата, распознающего устную речь, и построения алгоритма его работы приходится вычислять вероятность совпадения хотя бы одной из словоформ обрабатываемого текста с соответствующей лексемой, заданной в словаре автомата.

Предположим, что нужно определить вероятность того, что хотя бы одно из двух выбранных слов текста будет местоимением *он*.

Обозначим через A первое появление местоимения *он*, а через B — второе появление этого же местоимения. События A и B совместимы, поскольку можно извлечь одновременно слово *он* как из первого, так и из второго отрывков. Поэтому при решении нашей задачи необходимо воспользоваться формулой (5.16). Значение ста-

тистической вероятности согласно данным частотного словаря [39] равно 0,0099. Учитывая это, получаем

$$P(A + B) = 0{,}0099 + 0{,}0099 - 0{,}0099 \cdot 0{,}0099 \approx 0{,}020.$$

Теперь предположим, что распознающий автомат анализирует десять взятых наугад словоформ, и попробуем определить вероятность того, что хотя бы одна из этих словоформ окажется местоимением *он*. Для этого воспользуемся формулой (5.17), обозначив через A совпадение текстовой словоформы с местоимением *он*, а через C — появление в нашем опыте хотя бы одного *он*. Поскольку вероятность $P(A)$ для всех отрывков одинакова, на основании равенства (5.17) найдем

$$P(C) = 1 - (1 - 0{,}0099)^{10} \simeq 0{,}095.$$

Таким образом, вероятность получить хотя бы одно местоимение *он* при десятикратном извлечении словоформы из текста заметно выше вероятности получить его при однократном или двукратном извлечении.

8. Зависимые лингвистические события и условные вероятности. До сих пор мы имели дело с *независимыми* событиями, т. е. с такими событиями, вероятность появления которых не зависела от вероятности появления другого лингвистического события — эти вероятности называются *безусловными*. Однако языкознание сравнительно редко имеет дело с независимыми событиями. Обычно речь идет о *зависимых событиях* и *условных вероятностях*: даже вероятности появления букв, фонем, слогов, морфем и т. д. являются условными, так как зависят от позиции этих лингвистических объектов в слове, словосочетании и предложении. Например, как показывают табл. 5.3 и 5.4, буква *п* в начале русского слова имеет вероятность 0,207, а после начального *я* вероятность ее появления составляет всего 0,001.

Таблица 5.3

Распределение вероятностей первых букв русского слова

Буква	P	Буква	P	Буква	P
п	0,207	*я*	0,035	*е*	0,014
н	0,085	*з*	0,032	*в*	0,014
и	0,070	*т*	0,031	*л*	0,012
с	0,064	*ш*	0,030	*х*	0,010
о	0,052	*ф*	0,029	*ц*	0,008
в	0,051	*р*	0,021	*ж*	0,007
к	0,040	*б*	0,020	*щ*	0,003
м	0,038	*у*	0,020	*ю*	0,002
д	0,037	*г*	0,016	*й*	0,001
а	0,036	*ч*	0,015		

Таблица 5.4

Распределение вероятностей русских букв после цепочки Δ*я**

Буква	P	Буква	P	Буква	P
Пробел (Δ)	0,701	*г*	0,004	*ч*	0,001
в	0,157	*й*	0,003	*п*	0,001
з	0,036	*д*	0,002	*х*	0,001
р	0,031	*к*	0,002	*ш*	0,001
щ	0,016	*л*	0,001		
б	0,009	*м*	0,001		

* Таблицы 5.3 и 5.4 составлены путем обследования разговорных, беллетристических, научно-технических и публицистических текстов длиной примерно в 500 стр. (ок. 200 тыс. словоформ). Возможное с точки зрения лексических норм современного русского языка начальное двухбуквенное сочетание *яф* (ср. *яфетический*, *Яффа*) не встретилось и поэтому не учтено в таблице.

Рассмотрим соотношение независимых и зависимых событий, а также безусловных и условных вероятностей на примере искусственного лингвистического опыта.

Словоформа *мамам* (дательный падеж множественного числа от *мама*) составлена из букв разрезной азбуки. Карточки с буквами этого слова положены в урну. Производится испытание, состоящее в последовательном извлечении карточки с буквой и возвращении ее обратно в урну. Событием B считается извлечение буквы *м* в первом испытании (тогда $\bar{B}$ будет извлечение из урны не *м*, т. е. буквы *а*), событием A — извлечение буквы *а* во втором опыте (тогда $\bar{A}$ будет извлечение из урны не *а*, т. е. буквы *м*). В силу того, что вынутая в первый раз буква возвращается обратно в урну, перед вторым опытом количество букв в урне не изменяется. Поэтому вероятность события A является безусловной, поскольку она не зависит от того, была ли извлечена до этого из урны буква *м* (событие B) или буква *а* (событие $\bar{B}$), и остается равной 2/5. Безусловной является и вероятность события B. Если изменить условия опыта и не возвращать извлеченную букву обратно в урну, то вероятности получить при втором, третьем и т. д., извлечениях букву *а* или *м* будут существенно зависеть от того, какие буквы были извлечены перед этим из урны.

Пусть исходом первого извлечения была буква *м*; тогда вероятность вытащить при втором извлечении букву *а* составит 2/4 = 1/2. В том же случае, когда в результате первого опыта получена буква *а* (событие $\bar{B}$), вероятность вытащить второй раз букву *а* равна 1/4. Сходное положение возникает при определении вероятности появления буквы *м* (событие $\bar{A}$) во втором извлечении при условии, что в первый раз была получена буква *м* (событие B) или *а* (событие $\bar{B}$). Иными словами, события A и B являются зависимыми, а их вероятности — условными.

Условная вероятность события A при условии, что произошло событие B, обозначается $P(A/B)$. Так, в рассмотренном выше примере

$$P(A/B) = 1/2,\ P(\bar{A}/B) = 1/2,\ P(A/\bar{B}) = 1/4,\ P(\bar{A}/\bar{B}) = 3/4.$$

Условная вероятность события A, вычисленная при условии, что осуществилось несколько событий B_1, B_2, B_3, ..., обозначается $P(A/B_1B_2B_3\ldots)$.

Величина условной вероятности всегда заключена в том же отрезке, что и величина абсолютной вероятности, т. е.

$$0 \leqslant P(A/B_1B_2B_3\ldots) \leqslant 1.$$

4. Правило умножения вероятностей и вычисление вероятностей цепочек языковых элементов. Каждый текст или его часть можно рассматривать как совместное наступление некоторой линейной последовательности лингвистических событий — совместное появление цепочки словоформ, последовательности слогов, цепочек фонем или букв. Определение вероятностей появления этих цепочек опирается на теорему умножения вероятностей, согласно которой *вероятность совместного наступления двух событий равна произведению вероятности первого события на условную вероятность второго, вычисленную при условии, что первое событие имело место*:

$$P(AB) = P(A)\,P(B/A) \text{ или } P(AB) = P(B)\,P(A/B). \quad (5.18)$$

Из этой теоремы вытекают три важных следствия.

Следствие 1. *Если событие A независимо от B, то и событие B независимо от A.*

Для независимых событий теорема умножения вероятностей упрощается и принимает следующий вид: *вероятность произведения двух независимых случайных событий равна произведению их безусловных вероятностей*:

$$P(AB) = P(A)\,P(B). \quad (5.19)$$

Следствие 2. *Если события A и B независимы, то независимы также и пары событий $(\bar{A}, B)$, $(A, \bar{B})$, $(\bar{A}, \bar{B})$.*

Следствие 3. *Вероятность произведения зависимых событий A, B, C равна произведению вероятности одного из них на условную вероятность второго и на условную вероятность третьего, вычисленную при условии, что предыдущие оба события произошли:*

$$P(ABC) = P(A)\,P(B/A)\,P(C/AB). \quad (5.20)$$

Обобщая это следствие на n зависимых событий A_1, A_2, ..., A_n, получаем

$$P\left(\prod_{i=1}^{n} A_i\right) = P(A_1)\,P(A_2/A_1)\,P(A_3/A_1A_2)\ldots P\left(A_n/\prod_{i=1}^{n-1} A_i\right). \quad (5.21)$$

Выше в табл. 5.2—5.4 были приведены значения для условных и безусловных статистических вероятностей отдельных букв в текстах современного русского литературного языка. Используя эти таблицы и соотношения (5.18), (5.20), (5.21), можно рассчитать вероятности появления в письменных текстах современного русского языка различных двухбуквенных сочетаний.

Так, например, вероятность появления группы Δя равна

$$P(\Delta я) = P(\Delta)\,P(я/\Delta) = 0{,}174 \cdot 0{,}035 = 0{,}006 = 0{,}6\%.$$

Чтобы определить вероятность появления слова *я*, образуем трехсловное сочетание ΔяΔ, для которого

$$P(\Delta я\Delta) = P(\Delta)\,P(я/\Delta)\,P(\Delta/\Delta я) = 0{,}174 \cdot 0{,}035 \cdot 0{,}701 = 0{,}00427 = 0{,}4\%.$$

Для расчета вероятности появления морфемы *япон* формируем цепочку Δ*япон*; тогда

$$P(\Delta япон) = P(\Delta)\,P(я/\Delta)\,P(п/\Delta я)\,P(о/\Delta яп)\,P(н/\Delta япо).$$

«Словарь русского языка» под ред. С. И. Ожегова показывает, что после цепочки Δ*яп* единственно возможным продолжением будет диграмма *он**. Отсюда следует, что появления здесь букв *о* и *н* являются достоверными событиями, условная вероятность которых равна единице. Таким образом,

$$P(\Delta япон) = 0{,}174 \cdot 0{,}035 \cdot 0{,}001 \cdot 1 \cdot 1 = 0{,}00006 = 0{,}006\%.$$

5. Определение общей вероятности лингвистического события с помощью формулы полной вероятности. Если лингвистическое событие A может осуществиться вместе с одним и только одним из n несовместимых событий H_1, H_2, ..., H_n, называемых *гипотезами* и образующих полную группу событий, то для определения вероятности этого события используется формула полной вероятности:

$$P(A) = \sum_{i=1}^{n} P(H_i)\,P(A/H_i). \quad (5.22)$$

Таким образом, *вероятность события A равна сумме произведений вероятности каждой гипотезы на вероятность события при осуществлении этой гипотезы.*

Формула полной вероятности используется для вычисления общей вероятности лингвистического события при условии, что известны его вероятности в узко-тематических выборках.

Пусть, например, имеется английский научно-технический текст общей длиной в 400 тыс. словоупотреблений (около тысячи стандартных страниц). По тематике этот текст распадается на следующие четыре выборки разной длины:

* Мы опускаем в данном случае продолжение *япош(ка)*, вероятность появления которого с точки зрения норм русского литературного языка близка к нулю.

1) радиоэлектроника — 200 тыс. словоупотреблений (ок. 500 с.),
2) автомобилестроение — 100 тыс. словоупотреблений (ок. 250 с.),
3) судовые механизмы — 50 тыс. словоупотреблений (ок. 125 с.),
4) строительные материалы — 50 тыс. словоупотреблений (ок. 125 с.),

Словоформа are — множественное число настоящего времени глагола to be '*быть*' употреблена в 1-й выборке 1610, во 2-й — 1273, в 3-й — 469 и в 4-й — 346 раз. Аналогичным образом, словоформа machine 'машина, механизм' встретилась в 1-й выборке 98, во 2-й — 57, в 3-й — 9 и в 4-й — 19 раз. Эти данные взяты из работы [6, с. 80].

Необходимо определить вероятность того, что извлеченное наугад из нашего текста словоупотребление будет: а) словоформой are; б) словоформой machine.

Для этого будем считать появление словоформы are событием A, а появление machine — событием B. Рассмотрим также следующие четыре гипотезы: H_1 — принадлежность словоформы к текстам по радиоэлектронике, H_2 — к текстам по автомобилестроению, H_3 — к текстам по судовым механизмам, H_4 — к текстам по строительным материалам.

Считая доли указанных текстов в общей выборке вероятностями наших гипотез, находим:

$$P\ (H_1) = 200000/400000 = 0{,}5;\ P\ (H_2) = 100000/400000 = 0{,}25;$$
$$P\ (H_3) = P\ (H_4) = 50000/400000 = 0{,}125.$$

Условные вероятности события A (появление глагола are) при этих гипотезах соответственно равны:

$$P\ (A/H_1) = 1610/200000 = 0{,}008;\ P\ (A/H_2) = 1273/100000 = 0{,}012;$$
$$P\ (A/H_3) = 469/50000 = 0{,}009;\ P\ (A/H_4) = 346/50000 = 0{,}007.$$

Применяя формулу полной вероятности, определяем, что вероятность извлечь наугад из данного текста словоформу are равна

$$P\ (A) = P\ (H_1)\ P\ (A/H_1) + P\ (H_2)\ P\ (A/H_2) + P\ (H_3)\ P\ (A/H_3) + P\ (H_4)\ P\ (A/H_4) =$$
$$= 0{,}5 \cdot 0{,}008 + 0{,}25 \cdot 0{,}012 + 0{,}125 \cdot 0{,}009 + 0{,}125 \cdot 0{,}007 =$$
$$= 0{,}009 = 0{,}9\%.$$

Аналогичным образом находим условные вероятности события B:

$$P\ (B/H_1) = 98/200000 = 0{,}0005;\ P\ (B/H_2) = 57/100000 = 0{,}0006;$$
$$P\ (B/H_3) = 9/50000 = 0{,}0002;\ P\ (B/H_4) = 19/50000 = 0{,}0004.$$

По формуле полной вероятности получаем, что вероятность извлечь из данного текста словоформу machine составляет

$$P\ (B) = 0{,}5 \cdot 0{,}0005 + 0{,}25 \cdot 0{,}0006 + 0{,}125 \cdot 0{,}0002 + 0{,}125 \cdot 0{,}0004 = 0{,}000475 \approx 0{,}048\%.$$

6. Априорные и апостериорные вероятности. Измерение вероятностей лингвистических гипотез. До сих пор мы имели дело с так называемыми *априорными* вероятностями лингвистических событий. Эти априорные вероятности устанавливались интуитивно-эмпирически или теоретически до осуществления опыта, исходя из наших знаний об условиях σ этого опыта. Наши сведения о всех условиях опыта обычно неполны, поэтому априорные вероятности являются вероятностями некоторых лингвистических гипотез $H_1, H_2, \ldots, H_n$ об исходе эксперимента.

Получаемый при осуществлении этого эксперимента результат заставляет нас обычно произвести переоценку наших гипотез и придать им новые — *апостериорные* вероятности. Определение апостериорных вероятностей осуществляется, исходя из следующих соображений.

Пусть априорные вероятности гипотез до опыта соответственно равны $P\ (H_1), P\ (H_2), \ldots, P\ (H_n)$, а в результате опыта отмечено появление события A. Необходимо определить, как нужно изменить вероятности наших лингвистических гипотез в связи с осуществлением события A.

Согласно теореме умножения вероятностей для зависимых событий, вероятность совместного наступления события A и гипотезы H_i $(i = 1, 2, \ldots, n)$ составляет

$$P(AH_i) = P(A)\ P(H_i/A) = P(H_i)P(A/H_i). \qquad (5.23)$$

Отсюда следует, что

$$P(H_i/A) = \frac{P(H_i)\ P(A/H_i)}{P(A)}. \qquad (5.24)$$

Подставляя для $P(A)$ его выражение из формулы полной вероятности (5.22), имеем

$$P(H_i/A) = \frac{P(H_i)\ P(A/H_i)}{\sum_{j=1}^{n} P(H_j)\ P(A/H_j)}. \qquad (5.25)$$

Выражение (5.25) носит название **формулы Бейеса**, или формулы вероятностей гипотез.

Чтобы показать, как с помощью формулы Бейеса измеряются вероятности лингвистических гипотез, обратимся снова к извлечению из английского научно-технического текста словоформ are и machine (см. п. 5).

Предположим, что первая наугад взятая из английского научно-технического текста словоформа оказалась глаголом are (событие A). Необходимо найти вероятность того, что эта словоформа извлечена: а) из текста по радиоэлектронике (H_1); б) из текста по автомобилестроению (H_2); в) из текста по судовым механизмам (H_3); г) из текста по строительным материалам (H_4).

Вероятности того, что извлеченная словоформа принадлежит к той или иной тематической выборке, являются апостериорными вероятностями гипотез — точнее, условными вероятностями этих гипотез при условии, что произошло событие A. Используя соотношение (5.25), получим

$$P(H_1/A) =$$

$$= \frac{P(H_1)\,P(A/H_1)}{P(H_1)\,P(A/H_1)+P(H_2)\,P(A/H_2)+P(H_3)\,P(A/H_3)+P(H_4)\,P(A/H_4)} =$$

$$= \frac{0{,}5\cdot 0{,}008}{0{,}5\cdot 0{,}008+0{,}25\cdot 0{,}012+0{,}125\cdot 0{,}009+0{,}125\cdot 0{,}007} = 0{,}444.$$

Аналогичным образом

$$P\ (H_2/A) = 0{,}333,\ P\ (H_3/A) = 0{,}128,\ P\ (H_4/A) = 0{,}095.$$

Нетрудно заметить, что апостериорные вероятности гипотез о принадлежности словоформы аге к определенным подъязыкам, обусловленные появлением этой словоформы, заметно отличаются от их априорных вероятностей, полученных в п. 5.

Используя приведенные выше данные, определим апостериорные вероятности гипотез H_1, H_2, H_3, H_4 при условии, что из текста дважды извлекались две словоформы, причем оба раза этими словоформами оказался глагол аге. Эксперимент строился таким образом, что обе словоформы могли быть извлечены только из одной тематической выборки.

Двойное извлечение словоформы аге является сложным событием, представляющим собой произведение двух независимых событий. В связи с этим формула Бейеса для расчета апостериорных вероятностей наших гипотез принимает здесь следующий вид:

$$P(H_i/AA) = \frac{P(H_i)\,P(A/H_i\cdot A/H_i)}{\sum_{j=1}^{n} P(H_j)\,P(A/H_j\cdot A/H_j)} = \frac{P(H_i)\,[P(A/H_i)]^2}{\sum_{j=1}^{n} P(H_j)\,[P(A/H_j)]^2}. \quad (5.26)$$

Проведя несложные расчеты, получаем:

$$P\ (H_1/AA) = 0{,}369,\ P\ (H_2/AA) = 0{,}437,$$

$$P\ (H_3/AA) = 0{,}126,\ P\ (H_4\ /AA) = 0{,}068.$$

Нетрудно заметить, что здесь снова имеет место перераспределение вероятностей гипотез, причем на первое место выдвигается гипотеза о том, что обе словоформы принадлежат второй выборке. После однократного извлечения аге наибольшую вероятность имела гипотеза H_1.

На понятиях априорной и апостериорной вероятности строится теория решений, применение которой имеет большое будущее в инженерной лингвистике; эти понятия используются также при формулировке понятия логической вероятности, которая является отправным пунктом процедуры, измеряющей семантическую информацию в тексте [26].

§ 5. Информационные измерения в тексте

1. Энтропия как мера неопределенности лингвистического опыта. Мы уже несколько раз встречались с количественными оценками информации, содержащейся в тексте и слове. Однако отсутствие достаточных математических сведений не позволило нам дать строгое определение количества информации и описать процедуру ее вычисления. Теперь, когда эти необходимые сведения введены, можно дать более или менее последовательное определение как самого понятия количества информации, так и информационных измерений в тексте.

Количественные измерения информации можно осуществить, опираясь на два исходных понятия — в е р о я т н о с т и случайного лингвистического события и н е о п р е д е л е н н о с т и, присутствующей перед осуществлением опыта, результатом которого является указанное событие. Понятие вероятности подробно рассматривалось в предыдущих разделах, понятие же неопределенности и ее меры нуждается в специальном разъяснении.

Каждый лингвистический опыт связан с некоторой неопределенностью исхода. Если наш опыт состоит в последовательном угадывании букв неизвестного слова, то угадывание каждой буквы по мере движения от начала слова имеет свою неопределенность. Чем больше альтернатив при выборе возможного исхода опыта, тем больше его неопределенность; чем меньше таких альтернатив, тем меньше неопределенности в исходе опыта. Например, при последовательном угадывании букв слова *который* наибольшая неопределенность имеет место при выборе первой буквы (здесь вместо *к* может стоять любая буква русского алфавита, за исключением твердого и мягкого знаков), она будет значительно меньше в случае угадывания седьмой буквы при условии, что предыдущие шесть *которо* нам известны. В этой позиции возможны четыре альтернативы: либо *г*, либо *е*, либо *й*, либо *м*. Заметим, что и в первом, и во втором случае угадывание происходит в предположении, что все допустимые в той или иной позиции буквы равновероятны. Если же обратиться к угадыванию буквы, стоящей после цепочки *которог*, то исход этого угадывания полностью определен: с точки зрения норм русской письменной речи здесь может находиться только буква *о*. Неопределенность в этом случае равна нулю. Таким образом, между неопределенностью опыта и количеством равновероятных исходов обнаруживаются следующие две зависимости:

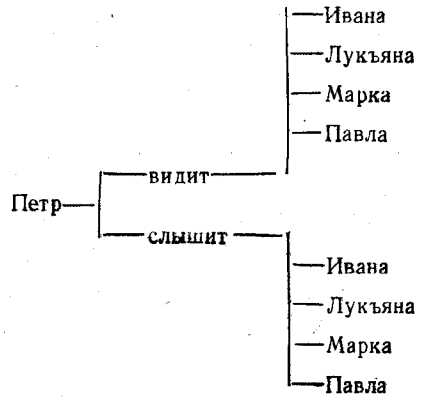
1) если число исходов $S = 1$, то неопределенность $f\ (S) = 0$;

2) если имеются два опыта, причем $S_1 > S_2$, то $f\ (S_1) > f\ (S_2)$.

Для того чтобы окончательно определить вид функции $f(S)$, характеризующей меру неопределенности, рассмотрим еще один лингвистический эксперимент.

Будем строить случайным образом трехсловное предложение. Пусть первая позиция занята именем собственным *Петр*. Вторую позицию нужно занять одной из двух глагольных словоформ *видит* или *слышит* ($S_1 = 2$), которые наугад извлекаются из урны.

Конечная позиция замещается одной из четырех словоформ — *Ивана*, *Лукъяна*, *Марка*, *Павла* ($S_2 = 4$), — также извлекаемых наугад из второй урны. Это построение можно изобразить в виде следующей схемы:



Неопределенность опыта, состоящего в выборе глагольной формы, равна $f(S_1) = f(2)$; неопределенность испытания, представляющего собой выбор имени собственного, характеризуется величиной $f(S_2) = f(4)$.

Теперь рассмотрим сложный опыт, заключающийся в комбинированном выборе из двух урн одного из $S_1 \cdot S_2 = 2 \cdot 4 = 8$ двухсловных продолжений для начальной словоформы *Петр*.

Неопределенность этого сложного опыта, являясь суммой неопределенностей двух простых опытов, характеризуется равенством

$$f(S_1 \circ S_2) = f(S_1) + f(S_2).$$

Последнее равенство представляет собой третью зависимость, характеризующую отношение между неопределенностью опыта и числом его равновероятных исходов.

Существует единственная функция аргумента S, отвечающая трем перечисленным выше условиям: 1) $f(1) = 0$; 2) если $S_1 > S_2$, то $f(S_1) > f(S_2)$; 3) $f(S_1 \circ S_2) = f(S_1) + f(S_2)$. Этой функцией является логарифмическая зависимость

$$H = \log S, \qquad (5.27)$$

с помощью которой мы будем оценивать меру неопределенности, или *энтропию*, опыта.

В лингвистических применениях энтропии, как правило, используются логарифмы при основании 2, в связи с чем выражение (5.27) принимает вид

$$H_0 = \log_2 S. \qquad (5.28)$$

Отсюда следует, что единицей измерения энтропии служит неопределенность, заключенная в опыте, содержащем два равновероятных исхода. Эта единица называется *двоичной единицей* (**дв.** ед.), или *битом*.

Вернемся к рассмотренному выше лингвистическому эксперименту с выбором продолжений для имени собственного *Петр*. Здесь неопределенность выбора глагольной формы языка

$$\log_2 2 = 1 \text{ (дв. ед.)},$$

а энтропия выбора имени собственного в третьей позиции составляет

$$\log_2 4 = 2 \text{ (дв. ед.)}.$$

Неопределенность же сложного опыта, состоящего в одновременном выборе сказуемого и прямого дополнения, должна составлять

$$\log_2 2 + \log_2 4 = 1 + 2 = 3 \text{ (дв. ед.)}.$$

Действительно,

$$\log_2 (2 \cdot 4) = \log_2 8 = 3 \text{ (дв. ед.)}.$$

2. Комбинаторный подход к определению количества информации. Введение понятия энтропии дает возможность проводить количественное измерение информации Действительно, в результате проведения опыта A мы получаем новые сведения, т. е. некоторую информацию. Одновременно знание исхода опыта снимает полностью или частично ту неопределенность, которая была до его осуществления. Естественно предположить, что снятая в результате опыта A энтропия количественно равна полученной информации, т. е.

$$H(A) = I(A). \qquad (5.29)$$

Из (5.28) и (5.29) следует, что количество информации, получаемое от испытания с множеством S равновероятных исходов, определяется равенством

$$I_0 = \log_2 S. \qquad (5.30)$$

Применительно к языковедческим задачам множество S называется *лингвистическим алфавитом*, а величины I_0 и H_0 — соответственно *информацией* и *энтропией алфавита*.

Число равновероятных исходов S определяется обычно путем исследования комбинаторики элементов и связей, характеризующих рассматриваемое лингвистическое явление. В связи с этим вся только что описанная методика представляет собой комбинаторный подход к определению количества информации [23, с. 7].

3. Измерение ограничений, накладываемых на употребление лингвистических единиц системой и нормой языка. Хотя комбинаторный подход дает, как правило, завышенные данные об энтропии и информации опыта, он может быть использован для полу-

чения приблизительных оценок тех ограничений, которые накладывают на употребление лингвистических единиц система и норма языка. Рассмотрим методику получения этих оценок на примере двухбуквенных сочетаний.

Исходя из соотношений (5.2) и (5.30), можно утверждать, что информация, получаемая от выбора такого двухбуквенного сочетания, которое строится средствами русского 32-буквенного алфавита при условии, что никаких ограничений на сочетаемость букв не накладывается и все двухбуквенные комбинации считаются равновероятными, составляет

$$I_0 = \log_2 \tilde{A}^2_{32} = \log_2 1024 = 10 \text{ (дв. ед.)}.$$

Если учесть ограничение, состоящее в том, что наши двухбуквенные сочетания не должны включать твердого и мягкого знака, то информация, содержащаяся в одной двухбуквенной комбинации, равна

$$I' = \log_2 \tilde{A}^2_{30} = \log_2 900 = 9{,}81 \text{ (дв. ед.)}.$$

Если же составить двухбуквенные комбинации из всех 32 букв русского алфавита, не допуская повторений букв, то, согласно (5.1), количество информации, получаемое от выбора одного буквосочетания, равно

$$I'' = \log_2 A^2_{32} = \log_2 992 = 9{,}95 \text{ (дв. ед.)}.$$

Легко заметить, что введение тех или иных ограничений на сочетаемость букв приводит к уменьшению информации, получаемой при выборе одного двухбуквенного сочетания. Эти ограничения, которые мы будем называть *структурными контекстными ограничениями*, можно количественно оценить с помощью разности

$$I_0 - I = K, \qquad (5.31)$$

где I_0 — информация алфавита или, иными словами, количество информации, которое извлекается из опыта при отсутствии каких-либо ограничений в комбинаторике лингвистических элементов и связей, I — информация, получаемая при учете интересующих нас ограничений, а K — контекстная обусловленность

Пользуясь выражением (5.31), нетрудно оценить величину структурных ограничений, накладываемых на алфавит русских двухбуквенных комбинаций. В первом случае эти ограничения составляют

$$K'_2 = \log_2 \tilde{A}^2_{32} - \log_2 \tilde{A}^2_{30} = 10 - 9{,}81 = 0{,}19 \text{ (дв. ед.)},$$

во втором случае

$$K''_2 = 10 - 9{,}95 = 0{,}05 \text{ (дв ед.)}.$$

Согласно данным словарей [22]; [35]; [39], в русском языке содержится около 250 двухбуквенных слов, из которых только

114 допущены нормой современного литературного языка. Отсюда следует, что лексическая система русского языка накладывает на образование осмысленного двухбуквенного слова структурные ограничения, равные

$$K \text{ (системы)} = \log_2 \tilde{A}^2_{32} - \log_2 250 = 10 - 7{,}96 = 2\,04 \text{ (дв ед.)}.$$

В то же время норма литературного языка дает дополнительные ограничения, составляющие

$$K \text{ (нормы)} = \log_2 250 - \log_2 114 = 1{,}73 \text{ (дв. ед.)}.$$

Комбинаторные измерения информации могут быть успешно применены для оценки «гибкости речи» т. е. при измерении разветвленности продолжения текста при заданном словаре и заданных правилах построения предложений.

4. Вероятностный подход к определению количества информации. При описании комбинаторного метода для вычисления количества информации и энтропии мы пользовались упрощающим допущением, согласно которому все исходы опыта считались равновероятными. Между тем при исследовании текста такая ситуация почти никогда не встречается. Норма языка и описываемая текстом ситуация приписывает каждому лингвистическому элементу определенную вероятность. Если лингвистическое испытание предусматривает разновероятные исходы, то, очевидно, энтропия такого опыта и получаемое от него количество информации будут отличаться от аналогичных величин, характеризующих опыт с равновероятностными исходами. Например, неопределенность опыта, состоящего в угадывании буквы, стоящей после цепочки *Аяпо*, гораздо меньше*, чем неопределенность опыта, состоящего в выборе равновероятных глагольных форм *видит* и *слышит*.

Переход от оценки неопределенности и информации опыта с равновероятными исходами к вычислению энтропии и информации испытания с неравновероятными исходами осуществляется на основе следующих соображений.

Опираясь на известные правила логарифмирования, перепишем выражение (5.30) в виде

$$I_0 = -\log_2 (1/S). \qquad (5.32)$$

Здесь величина $1/S$ есть не что иное, как вероятность p каждого исхода опыта. Предположим теперь, что исходы опыта неравновероятны и каждый исход i имеет свою вероятность p_i. Тогда индивидуальное количество информации, приносимое исходом i при его отдельном появлении, равно

$$I_i = -\log_2 p_i.$$

* Исходя из норм литературного языка, после цепочки *Аяпо* должна стоять буква *н* (ср. *японец, японка* и т. п.). Вероятность же появления буквы *ш* здесь очень мала (ср. редкое просторечное *япошка*); см. § 4, п. 4.

При многократном осуществлении опыта исход i будет происходить с вероятностью p_i. Поэтому среднее количество информации, приносимое исходом i при многократном осуществлении испытания, составит

$$\tilde{I}_i = -p_i \log_2 p_i.$$

Величина $\tilde{I}_i$ определяет тот вклад, который вносит исход i в общее количество информации, получаемой при многократном проведении опыта A. Что касается общей информации, то она, представляя собой сумму вкладов всех S возможных исходов, определяется равенством

$$I = -\sum_{i=1}^{S} p_i \log_2 p_i. \qquad (5.33)$$

Это равенство является исходной формулой вероятностного подхода к определению количества информации.

Величины I_i, $\tilde{I}_i$ и I имеют разную качественную интерпретацию и различное количественное значение. Это видно на таком лингвистическом примере. Предположим, нам известно, что буква *н* появляется после цепочки *Δяпо* с вероятностью $p_н = 0{,}999$, а буква *ш* встречается после этой цепочки один раз на тысячу случаев ($p_ш = 0{,}001$). Тогда, если данный исход опыта дает после цепочки *Δяпо* обычное *н*, мы получаем всего лишь

$$I_н = -\log_2 0{,}999 = 0{,}0014 \text{ (дв. ед.)}$$

информации. В то же время появление редкого *ш* приносит

$$I_ш = -\log_2 0{,}001 = 10 \text{ (дв. ед.)}$$

информации, т. е. в 7 тыс. раз больше, чем появление более частого *н*. Такое соотношение вполне соответствует здравому смыслу: тривиальный исход опыта всегда малоинтересен и малоинформативен, напротив, неожиданный результат всегда несет много информации.

Однако опыт дает редко неожиданный исход, поэтому вклад этого исхода в общую информацию опыта составляет всего лишь

$$I_ш = -0{,}001 \log_2 0{,}001 = 0{,}0100 \text{ (дв. ед.)}.$$

Это только в семь раз больше того вклада, который вносит в информацию опыта частый исход, информационный вес которого равен

$$I_н = -0{,}999 \log_2 0{,}999 = 0{,}0014 \text{ (дв. ед.)}.$$

Общее же количество информации, приносимое разновероятными исходами рассматриваемого опыта, составляет

$$I = -0{,}001 \log_2 0{,}001 - 0{,}999 \log_2 0{,}999 = 0{,}0114 \text{ (дв. ед)}.$$

Это заметно меньше информации, получаемой от опыта с двумя равновероятными исходами, где

$$I_0 = \log_2 2 = 1 \text{ (дв. ед.)}.$$

Информационные измерения, опирающиеся на вероятностный подход, могут быть осуществлены при условии, что для интересующего языковеда лингвистического опыта имеется полный набор вероятностей p (или оценивающих их частостей f) исходов этого опыта. Например, чтобы оценить информацию, которую несет в среднем одна буква русского алфавита (так называемая информация первого порядка I_1), необходимо обработать с помощью формулы (5.33) распределение (спектр) вероятностей букв в русских литературных текстах, показанный в табл. 5.2. Чтобы вычислить информацию, приходящуюся в среднем на одно слово или словоформу какого-либо языка или его разновидности, необходимо также обсчитать с помощью выражения (5.33) соответствующий частотный словарь [32 а, с. 179—261]; [39].

Однако информационные измерения, опирающиеся на обработку распределений безусловных вероятностей, имеют в языкознании ограниченное применение. Дело в том, что фонемы, графемы, слова и другие языковые единицы выступают в тексте в качестве зависимых лингвистических событий, обусловленных контекстом, а их вероятности являются условными (см. § 4, п. 3). Распределение последних вероятностей определяется тем положением, которое занимает данная лингвистическая единица в тексте. Так, например, распределение вероятностей русских букв в начале слова (см. табл. 5.3) сильно отличается от спектра из безусловных вероятностей (табл. 5.2) и совсем не похоже на распределения вероятностей букв, стоящих после цепочек *Δя* (табл. 5.4), *Δяп* или *Δяпо*.

Отсюда следует, что в большинстве случаев лингвистический опыт характеризуется не безусловной, а условной энтропией, определяющейся тем контекстным окружением, в котором находится данный участок текста. Так, например, выбор начальной буквы в русском слове имеет иную неопределенность, чем энтропия выбора буквы после цепочки *Δяпо*, т. е.

$$H \text{ (буквы/}\Delta) \neq H \text{ (буквы/ } \Delta япо),$$

и т. п. Само собой разумеется, что распределение вероятностей исходов и неопределенность опыта могут быть обусловлены не только предшествующим, но и последующим контекстом.

Что же касается информации, которая извлекается из данного участка текста, то она равна энтропии, характеризующей этот участок.

Рассмотрим теперь в деталях методику вычисления информации, получаемой от некоторого лингвистического опыта L, имеющего S исходов и осуществляющегося в n-м участке текста при условии, что стоящая перед этим участком цепочка лингвистических элементов b^{n-1} известна. Цепочка b^{n-1} рассматривается в качестве случайного события, принимающего частный вид i. Появ-

ление того или иного элемента в позиции n также рассматривается в качестве случайной величины, принимающей значение j_k ($1 \leqslant k \leqslant S$). Для каждого значения i, которое может принять b^{n-1} имеется условная вероятность $p\,(j_{i,k}/b_i^{n-1})$ того, что L_n примет значение j_k.

Средняя условная энтропия H_n, количественно равная информации I_n, получается в результате осреднения энтропии, подсчитанной по всем значениям b_i^{n-1} с весами, соответствующими вероятностям цепочки b^{n-1}. Таким образом, имеем

$$H_n = I_n = -\sum_{b_i^{n-1}} p\,(b_i^{n-1}) \sum_{k=1}^{S} p\,(j_{i,k}/b_i^{n-1}) \log p\,(j_{i,k}/b_i^{n-1}). \quad (5.34)$$

Равенство (5.34) показывает, какова в среднем мера неопределенности и количество информации от выбора лингвистического элемента в позиции n, когда известна цепочка b^{n-1}.

Если взаимосвязи элементов распространяются как угодно далеко, то энтропия (информация) на один лингвистический элемент составляет

$$H_\infty = I_\infty = \lim_{n\to\infty} H_n \text{ (дв. ед.)}. \quad (5.35)$$

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

$$H_n = I_n = H\,(j/b_i^{n-1}) = H\,(b_i^n) - H\,(b_i^{n-1}) \text{ (дв. ед.)}. \quad (5.36)$$

Формула (5.36) используется для определения информации, которую несет буква при условиях, что: предшествующая ей буква известна (информация второго порядка I_{II}); известны две предшествующие буквы (информация третьего порядка I_{III}) и т. д. [23, с. 10].

Объем работы при вычислении I_I, I_{II}, I_{III}, I_{IV} для буквенного и фонемного алфавита сравнительно велик. Существуют также реалистичные способы оценки I_I для слов, словоформ и даже словосочетаний. Однако число комбинаций из пяти, шести и т. д. букв (фонем), не говоря уже о комбинациях из четырех, пяти и т. д. слов (слогов, морфем), растет так быстро, что для определения их условной энтропии требуется колоссальная счетная работа.

Поэтому необходимо искать косвенные методы, с помощью которых можно было бы достаточно быстро определить энтропию и информацию в различных частях текста. Этому требованию отвечает эксперимент по предсказанию букв неизвестного текста, опирающийся на субъективные оценки условных вероятностей этих букв (ср. § 3, п. 1) со стороны угадчика.

В основе этого метода лежит предположение, что в сознании носителя языка заложены достаточно полные сведения о вероятностных характеристиках нормы языка. На основании его лингвистического опыта в сознании угадчика формируется степень убежденности по поводу того, какие лингвистические единицы встречаются чаще, а какие — реже, и хотя «ранжирование» лингвистических элементов в сознании разных носителей языка может быть неодинаковым, оно приближается при достаточно хорошем знании языка к некоторой оптимальной схеме.

Испытуемый или коллектив испытуемых угадывает текст, опираясь каждый раз на ранжированный спектр лингвистических единиц. Поэтому всю совокупность предсказаний можно рассматривать как некоторую систему лингвистических реакций, более или менее полно отражающую статистические процессы образования текста, за которыми стоят вероятностные свойства нормы языка.

Если осуществляется коллективное угадывание, то результаты его обрабатываются по формуле (5.33), причем вероятность определяется здесь из равенства $p_i = n_i/N$, где аргументом n_i является число угадчиков, предложивших ту или иную букву (независимо от того, правильна она или нет), а N есть общее число угадчиков.

При индивидуальном угадывании по полной программе используются две формулы: по одной определяется верхняя граница интервала, в котором находится истинное значение информации:

$$\bar{I} = (1-q_0)\log(1-q_0) - \sum_{i=1}^{S} q_i \log q_i, \quad (5.37)$$

а по другой — нижняя граница:

$$\underline{I} = \sum_{i=2}^{S} (q_i - q_{i+1})\, i \log i. \quad (5.38)$$

Величины q_i указывают на вероятность правильно угадать букву с i-й попытки, а q_0 есть вероятность достоверного продолжения.

Результаты сокращенной программы угадывания обрабатываются по формуле

$$\bar{I}' = H_{III}\ (1 - q_0 - q_1) + (1 - q_0)\log(1 - q_0) - \\ - q_1 \log q_1 - (1 - q_0 - q_1)\log(1 - q_0 - q_1), \quad (5.39)$$

где H_{III} есть неопределенность (энтропия) третьего порядка, значения остальных символов те же, что и в выражениях (5.37) и (5.38). Подробнее о психо-лингвистической стороне угадывания см. [23, с. 12—14].

Кроме того, делаются попытки выработать новые приемы расчета информации [23, с. 53—55]; [41, с. 258], в том числе и такие, которые не предусматривают обращения к вероятностному или теоретико-множественному подходу [53, с. 131—150].

5. Синтактическая информация и особенности ее распределения в тексте и слове. Полученные при алгоритмическом и вероят-

ностном подходе информационные величины не измеряют семантики сообщения. Они ничего не говорят и о том, как интерпретирует содержание текста его получатель (прагматика сообщения).

Получаемые в результате психо-лингвистического или просто комбинаторно-статистического эксперимента величины являются синтактической (по иной терминологии — статистической или селективной) информацией и служат количественной мерой структурно-статистического разнообразия и свободы выбора лингвистического варианта. Это разнообразие и свобода выбора задаются в каждом участке текста системой и нормой языка. Вместе с тем результаты нашего эксперимента отражают и то разнообразие, которое существует внутри самой системы и нормы языка (ср. в этом смысле оценки энтропии распределений длин слогов и морфем, которые будут приведены в гл. 6).

Наблюдения над распределением синтактической информации в тексте и в отдельно взятом слове, о которых уже говорилось выше (см. гл. 1, § 8; гл. 2, § 4; гл. 4, § 2 и 3), раскрывают некоторые важные особенности функционирования языка в речи.

В частности, выясняется, что текст дает квантовое распределение информации. Очевидно, письменная и устная речь воспринимается и перерабатывается нашей памятью не непрерывно, а путем ритмической отдачи накопленных порций информаций. Периодичность обнаруживается и в информационном построении длинных слов. Похоже, что в качестве кванта синтактической информации выступает морфема — элементарная смыслонесущая единица текста (см. гл. 1, § 9, п. 3).

Выше было показано, что в письменном тексте основная часть информации размещается в начале слова. Концы слов, а у длинных слов и средние участки оказываются избыточными. Такое распределение объясняет не только механизм аббревиации, но проясняет также вопрос об установлении границы слова в тексте. Интерпретант текста (читатель или слушатель), принимая решение о правой границе слова, не опирается на граничные комбинации графем или фонем, а прогнозирует положение конца слова, исходя из его информационно нагруженного начала.

Неравномерностью распределения синтактической информации в слове можно объяснить и особенности развития некоторых языков. Сюда относится фонетическое выветривание середин и особенно концов слов, приводящее к ослаблению и разрушению флексий, а одновременно значительная сопротивляемость фонетическим изменениям начал слов и препозитивных служебных форм [27]; [37]; [58].

6. Контекстная обусловленность и избыточность текста. Если синтактическая информация выступает в качестве меры свободы выбора лингвистического варианта в данном участке текста, то контекстная обусловленность используется в качестве меры тех ограничений, которые накладывают на данный участок текста система и норма языка.

Контекстная обусловленность задается уже известным нам выражением (5.31), которое применительно к синтактической информации, полученной вероятностным путем, имеет вид

$$K = I_0 - I_n. \tag{5.40}$$

Значения синтактической информации и контекстной обусловленности сильно зависят от объема лингвистического алфавита (числа букв, фонем, слогов и т. п.), использующегося в данном языке. Поэтому при сравнении разных языков удобнее пользоваться взвешенной величиной контекстной обусловленности, которая не зависит от длины алфавита. Эта взвешенная величина называется *избыточностью* в данном участке текста:

$$R_n = (I_0 - I_n)/I_0. \tag{5.41}$$

Общая избыточность текстов данного языка или его разновидности определяется из равенства

$$R = (I_0 - I_\infty)/I_0. \tag{5.42}$$

Значения избыточности, полученные для семи языков и их разновидностей из экспериментальных значений $\underline{I}_\infty$ и $\overline{I}_\infty$ [см. формулы (5.37) и (5.38)], показаны в табл. 5.5. Легко заметить, что все исследованные языки в целом имеют близкие значения избыточности. Вместе с тем явные расхождения обнаруживаются между значениями

Таблица 5.5

Избыточность (в %) некоторых индоевропейских и тюркских языков и их разновидностей по нижней ($\underline{R}$) и верхней ($\overline{R}$) границам

Языки / Разновидности	Русский		Польский		Английский		Немецкий	
	$\underline{R}$	$\overline{R}$	$\underline{R}$	$\overline{R}$	$\underline{R}$	$\overline{R}$	$\underline{R}$	$\overline{R}$
Разговорная речь	72,0	83,4	76,3	86,3	69,4	81,2	73,9	84,4
Беллетристика	76,3	86,0	74,5	83,6	77,1	86,5	71,4	82,5
Деловая речь (научно-технические и публицистические тексты)	83,4	90,1	83,6	89,5	82,9	92,1	79,6	88,2
Язык в целом	72,1	83,6	74,7	85,0	71,9	84,5	71,4	85,1

Продолжение табл. 5.5

Языки / Разновидности	Французский		Румынский		Казахский	
	$\underline{R}$	$\overline{R}$	$\underline{R}$	$\overline{R}$	$\underline{R}$	$\overline{R}$
Разговорная речь	68,5	82,9	74,2	85,4	76,5	85,0
Беллетристика	71,0	83,6	73,8	83,8	75,0	85,0
Деловая речь (научно-технические и публицистические тексты)	83,9	90,4	74,4	85,7	78,5	88,0
Язык в целом	70,6	83,4	72,1	85,0	71,9	85,0

избыточности, характеризующими отдельные разновидности языка. В частности, высокую избыточность показывает в некоторых языках деловая речь. Вопрос о существенности расхождений в численных значениях избыточности по отдельным стилям мы рассмотрим в гл. 9.

7. Измерение смысловой информации в тексте. Количественные оценки смысловой информации, содержащейся в тексте и в образующих его словах и словосочетаниях, можно получить, опираясь на значения синтактической информации и пользуясь идеей контекстной обусловленности.

В ходе эксперимента по угадыванию букв неизвестного текста было замечено, что свои гипотезы о наиболее вероятных продолжениях текста испытуемые строят, исходя из двух типов комбинаторных ограничений: комбинаторики фигур (букв и слогов) и комбинаторики знаков (морфем, слов, словосочетаний).

Эксперимент показывает, что уже на 4-м или 5-м буквенных шагах текста комбинаторика букв и слогов подавляется ограничениями, связанными с сочетаемостью морфем и слов. Затем по мере развертывания текста на комбинаторику слов напластовываются ограничения в сочетаемости словосочетаний и предложений, затем появляются ограничения, связанные с комбинаторикой параграфов, глав, частей книги или статьи.

Таким образом, при угадывании букв, находящихся на достаточном удалении от начала текста, испытуемый опирается не на статистическую комбинаторику букв и слогов, а на смысловое (лексико-грамматическое) построение текста. Поэтому если информация, извлеченная из начального участка экспериментального текста, выступает как количественная оценка дистрибуции (распределения) и статистики букв, то синтактическая информация, которая получается с удаленных от начала текста участков, служит отражением смысловой (семантико-прагматической) информации.

Эти соображения позволяют предложить некоторые приемы для количественной оценки смысловой информации, содержащейся в тексте и его сегментах.

Начнем с оценки смысловой лексико-грамматической информации, содержащейся в отдельном слове.

Пусть имеется текст, представляющий собой цепочку слов

$$w_1, w_2, w_3, \ldots, w_k, \ldots,$$

и мы хотим оценить количество информации, содержащееся в слове w_1.

Для решения этой задачи проведем коллективное угадывание сегмента текста $w_2 \div w_k$. Первый раз коллективу сообщено слово w_1, стоящее перед контрольным сегментом. Второй раз угадывание начинается прямо со слова w_2 (само собой разумеется, что между обоими угадываниями должно пройти достаточно времени, чтобы испытуемые забыли содержание текста; либо угадывания должны быть проведены в двух разных, но идентичных коллективах).

Естественно, что оба угадывания дадут разные результаты.

Получаемая в первом случае от контрольного сегмента информация

$$I\ (w_2 \div w_k) = H\ (w_2 \div w_k)\ \text{(дв. ед.)} \qquad (5.43)$$

будет больше информации

$$I\ (w_2 \div w_k/w_1) = H\ (w_2 \div w_k/w_1)\ \text{(дв. ед.)}, \qquad (5.44)$$

полученной при условии, что испытуемым было известно слово w_1. Разность

$$И\ (w_1) = I\ (w_2 \div w_k) - I\ (w_2 \div w_k/w_1)\ \text{(дв. ед.)} \qquad (5.45)$$

представляет собой количественную оценку той смысловой (семантико-прагматической) информации, которая содержится в слове w_1. Именно эта информация уменьшила неопределенность контрольного сегмента [ср. неравенство $H\ (w_2 \div w_k/w_1) < H\ (w_2 \div w_k)$] и облегчила второе угадывание.

Результаты первых пробных экспериментов по количественной оценке значений отдельных слов в четырех языках показаны в табл. 5.6. Для того чтобы проводить строгое сравнение оценок информаций по разным словам внутри одного языка и по одному слову для разных языков, необходимо иметь средние оценки смысловой информации. Такие данные можно получить, усредняя оценки, снятые с большого количества разных контекстов. Однако даже предварительные результаты показывают, что в аналитических языках (французском и болгарском) слово несет меньше смысловой информации, чем это имеет место в синтетических (флективном русском и агглютинирующем эстонском) языках. Эти наблюдения согласуются с традиционными предположениями об информационной нагруженности слов в указанных языках.

Теперь попробуем оценить ту лексико-грамматическую информацию, которую содержит контекст. Для этого с помощью приемов, описанных в работах [23, с. 69—73, 81—82] и [26], определим количество синтактической информации, приходящееся на слово средней длины l, взятое вне контекста, т. е. $\overline{I}\ (w_1)$, и в контексте, т. е. $\overline{I}\ (w_k)$. Кроме того, с помощью выражения

$$I\ (w_0) = lU_0\ \text{(дв. ед.)}$$

определим то количество информации, которое может нести цепочка длиной в l символов (букв) при условии, что все символы алфавита данного языка равновероятны и обладают неограниченными возможностями сочетаемости.

Тогда общая сумма контекстных ограничений, накладывающихся на слово в контексте, для верхней границы информации составит

$$\overline{K}\ (w_k) = \overline{I}\ (w_0) - \overline{I}\ (w_k)\ \text{(дв. ед.)}. \qquad (5.46)$$

Таблица 5.6

Количественные оценки семантической информации, содержащейся в некоторых русских, болгарских, французских и эстонских словах

Русский				Болгарский				Эстонский				Русский			
Слова	$H(w_0 \div w_k)$	$H(w_1 \div w_k/w_1)$	$И(w_1)$	Слова	$H(w_2 \div w_k)$	$H(w_2 \div w_k/w_1)$	$И(w_1)$	Слова	$H(w_2 \div w_k)$	$H(w_2 \div w_k/w_1)$	$И(w_1)$	Слова	$И(w_1)$
корреспондент	50,70	34,82	15,88	корес-пондентът	38,21	26,06	12,15	—	83,77	65,48	18,29	класс	18,33
печать	47,44	31,70	15,74	печатът	30,50	27,65	2,85	—	79,85	59,22	20,63	государство	17,66
									52,18	35,04	17,14	эксплуатация	19,84
								klass	72,85	55,75	17,10	правительство	13,36
								riik				народ	
								ekspluatee-rimine	76,33	58,69	17,64	президент	8,55
								valitsus					8,44
Французский								rahvas					

Слова							
la classe	39,25	30,00	9,25	—	—	—	—
l' état	49,21	42,16	7,05	—	—	—	—
l' exploitation	43,82	34,27	9,55	—	—	—	—
le gouvernement	38,76	33,34	5,42	—	—	—	—
le peuple	44,30	37,45	6,85	президентът	58,19	50,98	7,21
le président	42,75	35,25	7,50				
le président	42,57	37,44	5,13	председателят	32,35	25,63	6,72
le regime	35,91	29,24	6,67	режимът	45,70	34,40	11,30

Контекстные ограничения $\overline{K}(w_k)$ включают информацию, которая характеризует вероятностную дистрибуцию букв и слогов (см. выше), а также ту синтактическую информацию, которая оценивает среднюю величину смысловой информации, содержащейся в лексико-грамматических связях контекста $w_1 \div w_{k-1}$, предшествующего слову w_k. Эта лексико-грамматическая информация, определяющая предсказуемость слова w_k, может быть получена из равенства

$$\overline{K}(w_k^T) = \overline{I}(w_1) - \overline{I}(w_k) \text{ (дв. ед.)}, \qquad (5.47)$$

поскольку, уже начиная со второго слова, угадывание текста осуществляется на основе смысловой информации.

Доля лексико-грамматических связей контекста относительно общей суммы ограничений, накладывающихся на текстовое слово, т. е. выражение

$$\overline{A} = \frac{\overline{K}(w_k^T)}{\overline{K}(w_k)} \cdot 100\ \%, \qquad (5.48)$$

может служить оценкой степени аналитичности языка. Значения $\overline{A}$ и другие контекстные оценки относительно четырех европейских языков приведены в табл. 5.7.

Таблица 5.7

Контекстные оценки и процент аналитизма в некоторых индоевропейских языках

Информационные величины (в дв. ед.) \ Языки	Русский	Английский	Французский	Румынский
$I(w_0)$	31,85	25,97	25,23	27,07
$\overline{I}(w_1)$	13,50	12,59	10,88	13,02
$\overline{I}(w_k)$	8,15	5,41	6,46	7,77
$\overline{K}(w_k^T)$	5,35	7,18	4,42	5,25
$\overline{K}(w_k)$	23,70	20,56	18,77	19,30
$\overline{A}$ (%)	22,57	35,00	23,55	27,01

То, что значение коэффициента аналитичности $\overline{A}$ для английского языка в полтора раза превосходит его значение в русском языке, хорошо согласуется с нашими представлениями о соотношении аналитизма и синтетизма в этих языках. Несколько неожиданным кажется малая величина коэффициента $\overline{A}$ в аналитическом

французском языке. Этот парадокс следует отнести за счет флективности таких частотных служебных слов, как артикль, неударные личные местоимения 3-го лица, притяжательные местоимения, которые дают строю французского языка флективно-аналитический характер.

Рассматриваемая лексико-грамматическая информация является по своей природе семантико-прагматической информацией. При этом степень и качество прагматизма зависит от организации эксперимента. При индивидуальном эксперименте речь идет о прагматизме угадчика, коллективное угадывание отражает прагматизм угадывающего коллектива. Последнее обстоятельство дает возможность использовать эксперимент по угадыванию для количественных оценок субъективного и коллективного (профессионального, социального, возрастного) восприятия смысловой информации, содержащейся в тексте.

ГЛАВА 6

ВЕРОЯТНОСТНОЕ МОДЕЛИРОВАНИЕ ПОРОЖДЕНИЯ ТЕКСТА И СОСТАВЛЯЮЩИХ ЕГО ЕДИНИЦ

§ 1. Повторение независимых испытаний в тексте

Как уже говорилось (см. гл. 5), при исследовании механизмов порождения текста результаты отдельного лингвистического испытания не представляют большого интереса. Изучение взаимодействия системы, нормы и ситуации эксплицируется с помощью моделей теории вероятностей, предусматривающих осуществление массового эксперимента, при котором одно и то же лингвистическое испытание повторяется много раз. Эти повторяющиеся испытания образуют серии, в каждой из которых интересующее нас событие появляется или не появляется определенное число раз.

Так, например, при вероятностном исследовании норм употребления слова *море* в произведениях А. С. Пушкина нас не будет интересовать, употреблено ли это слово, скажем, в первом предложении на 100-й странице первого тома Академического издания сочинений поэта. Зато нам важно будет дать прогноз общего числа появлений этого слова в определенном числе (серии) предложений, составляющих, например, текст «Евгения Онегина», либо образующих некоторую выборочную последовательность из писем поэта и т. д.

Выбор той или иной модели описания текста зависит от построения вероятностно-лингвистического испытания и, в частности, от того, как организовано извлечение из текста отдельных его единиц.

1. Повторная и бесповторная выборки. Рассмотрим следующий элементарный пример. Пусть из текста взято N фонем, среди которых n гласных и m согласных, и каждая из фонем записана на отдельную карточку; карточки положены в урну и перемешаны. Испытания, состоящие в извлечении из урны одной карточки, могут осуществляться по двум схемам.

Согласно условиям первой схемы каждая извлеченная карточка возвращается в урну, после того как в протоколе фиксируется результат каждого испытания. При каждом последующем испытании вероятности появления гласной или согласной остаются неизменными. (Эти вероятности соответственно равны n/N и m/N.) Вероятностно-лингвистический эксперимент, оперирующий с последствиями взаимно независимых испытаний, в каждом из которых лингвистические события сохраняют свои безусловные вероятности, называется *повторной выборкой*.

При реализации второй схемы извлеченные из урны карточки не возвращаются. Вероятность появления гласной и согласной при каждом последующем испытании зависит от результатов предшествующих извлечений. Таким образом, мы имеем здесь дело с зависимыми испытаниями, а вероятность исхода каж-

дого из испытаний является условной (ср. гл. 5, § 4, п. 3). Эксперимент, оперирующий с последовательностью зависимых испытаний, в каждом из которых исходы имеют условные вероятности, называется *бесповторной (безвозвратной) выборкой*.

Реальный вероятностно-лингвистический эксперимент может быть осуществлен как с помощью повторной, так и с помощью бесповторной выборки.

При повторной выборке подвергшиеся испытанию лингвистические единицы должны каждый раз как бы возвращаться в текст. Рассмотрим организацию повторной выборки на следующем примере.

Предположим, что необходимо определить статистическую вероятность имен существительных или отдельных словоформ (скажем, *и*, *при*, *можно*, *напряжение*) в русских текстах по вычислительной технике. Для этого выберем массив текста из книг: А. И. Китов и Н. А. Криницкий. Электронные цифровые машины и программирование (М., 1959, с. 1—566) и А. А. Паперное. Логические основы цифровых машин и программирование (М., 1965, с. 7—440). Во второй книге страницы перенумеруем таким образом, что они образуют последовательность от 567-й до 1000-й страницы. Этим способом формируется массив в 1000 страниц, содержащий около 400 тыс. словоупотреблений.

Поиск лингвистических единиц осуществляется здесь следующим образом. Из таблицы случайных чисел последовательно выбираются шестиразрядные числа. Эти числа служат адресами тех словоупотреблений, которые мы будем сопоставлять при каждом испытании с интересующей нас единицей. Первые три цифры указывают на страницу, следующие две — на строку, а последняя — на номер словоупотребления в строке.

В тех случаях, когда то или иное случайное число указывает на несуществующую страницу, строку или словоупотребление, адрес считается недействительным и выборка осуществляется по следующему числу.

Если выбранное таким образом словоупотребление оказывается интересующей нас лингвистической единицей, то мы имеем дело с благоприятным исходом испытания. В противном случае имеет место неблагоприятный исход. Сумма благоприятных исходов, деленная на общее число испытаний, даст статистическую вероятность интересующей нас лингвистической единицы.

Возьмем теперь первые двадцать пять чисел из таблицы случайных чисел: 857454, 457562, 499988, 762760, 431557, 698780, 038799, 558121, 653187, 573553, 609209, 179138, 974652, 011813, 098638, 805797, 516103, 296103, 149471, 815377, 070381, 692830, 696116, 203055, 350356.

При этом реальные адреса дают нам 8, 9, 11, 18, 19, 21, 23, 24 и 25-е случайные числа, по которым из текста соответственно выбираются следующие словоупотребления: *таким*, *делению*, *бы*, l_1, *остается*, *зависимости*, *наименование*, *система*, *оба*. Остальные адреса являются недействительными.

Таким образом, нам удалось выбрать четыре существительных, вместе с тем мы не встретили ни одной из исследуемых словоформ.

При обследовании высокочастотных единиц, требующих сравнительно небольшой серии испытаний (к таким единицам относятся части речи и члены предложений, знаки препинания, классы букв и фонем), желательно в целях математической строгости применять схему независимых испытаний (повторную выборку).

Если же речь идет об определении вероятности таких редких единиц, как словосочетания, словоформы, фонемы и их сочетания, то осуществление повторной выборки вручную оказывается неосуществимой задачей из-за большого объема работы, связанного с громоздкой процедурой извлечения этих единиц из текста*. Поэтому здесь приходится применять бесповторную выборку. Хотя бесповторная выборка представляет собой последовательность зависимых испытаний, математическая обработка ее результатов производится обычно исходя из схемы независимых испытаний. Ниже будет показано (см. п. 6), что при большом объеме исследуемой лингвистической совокупности это нарушение математической строгости не приводит к сколько-либо заметным искажениям конечных результатов.

2. Три схемы независимых лингвистических испытаний. Квантитативное языкознание широко использует метод серийного наблюдения. Сущность его заключается в том, что лингвистические единицы выбираются из текста группами фиксированной длины: например, по десять фонем, по сто предложений или словоформ и т. п. Лингвистические единицы, составляющие серию, не обязательно должны находиться в тексте рядом друг с другом, они могут извлекаться и через определенный интервал.

При решении многих теоретических и инженерно-лингвистических задач оказывается необходимым знать вероятность появления того или иного числа интересующих исследователя лингвистических единиц в серии.

Если образующие серию лингвистические испытания рассматриваются как независимые, то мы можем осуществлять необходимое прогнозирование с помощью разработанных в теории вероятностей трех систем независимых испытаний: простой, полиномиальной и пуассоновской.

Простая схема предусматривает только два исхода опыта: появление или непоявление признака A. Примером этой схемы является повторная выборка из текста согласных (A) и гласных ($\bar{A}$) фонем (см. п. 1).

В *полиномиальной схеме* испытание дает не два, а несколько исходов. По этой схеме осуществляется, например, эксперимент, заключающийся в выборе из текста графем трех видов: букв, знаков препинания и пробелов.

---

* В настоящее время рассматривается вопрос о создании машинного алгоритма повторной выборки редких лингвистических единиц из больших массивов текста. Адреса выборки генерируются случайным образом самой ЭВМ.

В *пуассоновской схеме* независимые испытания осуществляются относительно нескольких совокупностей (подъязыков, стилей, тематик), в каждой из которых данный признак имеет разную вероятность. Поэтому вероятность лингвистического исхода меняется в зависимости от того, относительно какого подъязыка или тематики производится опыт.

Математическая модель, по которой осуществляется прогнозирование результатов простой схемы испытаний, является исходной при построении других вероятностных моделей, в том числе и тех, которые широко используются в квантитативной лингвистике. Поэтому мы особенно детально рассмотрим математическую модель простой схемы независимых испытаний.

3. Простая схема независимых испытаний. Формула Бернулли. Предположим, что в некотором тексте длиной в n фонем имеется m согласных и $n - m$ гласных. По схеме повторной выборки производится N независимых испытаний, заключающихся в последовательном случайном извлечении фонемы из текста. Требуется определить вероятность события, состоящего в том, что среди извлеченных N фонем ровно x окажутся согласными, причем порядок следования гласной и согласной фонем безразличен

Считая появление согласной событием A, а гласной — событием $\bar{A}$, определим вероятности появления гласной и согласной. Согласно классическому определению вероятности, имеем

$$P(A) = m/n = p, \quad P(\bar{A}) = (n-m)/n = q.$$

Теперь найдем вероятность того, что при N независимых испытаниях событие A появится ровно x раз, если вероятность появления этого события при каждом отдельном испытании постоянна и равна p.

Для этого составим всевозможные схемы, которые представляют разнообразную последовательность из появления x раз события A и $N - x$ раз его непоявления, т. е.

$$\underbrace{A\,A \ldots A}_{x \text{ раз}}\ \underbrace{\bar{A} \ldots \bar{A}}_{N-x \text{ раз}}.$$

По теореме умножения вероятность появления каждой схемы составляет $p^x q^{N-x}$, а число таких схем равно числу сочетаний из N элементов по x, т. е. C_N^x. Отсюда следует, что вероятность появления события A ровно x раз в серии N независимых испытаний составляет

$$P_N(x) = C_N^x\, p^x q^{N-x} = \frac{N!}{x!\,(N-x)!}\, p^x q^{N-x}, \qquad (6.1)$$

где $p + q = 1$. Заметим, также, что вероятности (6.1) равны соответствующим членам разложения по формуле бинома выражения $(q + p)^N$.

С помощью выражения (6.1), носящего название **формулы Бернулли**, и осуществляется вероятностное прогнозирование результатов в простой схеме независимых испытаний.

Все возможные несовместимые между собой исходы N опытов состоят в появлении 0, 1, 2, ..., N раз события A. Поэтому сумма величин (6.1), представляющих собой отдельные значения вероятностей при $x = 0, 1, 2, \ldots, N$, равна единице:

$$\sum_{x=0}^{N} P_N(x) = \sum_{x=0}^{N} C_N^x\, p^x q^{N-x} = (q+p)^N = 1.$$

Распределение вероятностей $P_N(x) = C_N^x p^x q^{N-x}$ при $x = 0, 1, 2, \ldots, N$, называемое *биномиальным распределением* (биномиальным законом распределения) вероятностей, можно записать в виде табл. 6.1.

Таблица 6.1

x	0	1	2	...
$P_N(x)$	$C_N^0 p^0 q^N = q^N$	$C_N^1 pq^{N-1} = Npq^{N-1}$	$C_N^2 p^2 q^{N-2} = \frac{N(N-1)}{2} p^2 q^{N-2}$	...

Продолжение табл. 6.1

x	...	$N-2$	$N-1$	N
$C_N^x p^x q^{N-x}$	...	$C_N^{N-2} p^{N-2} q^2 = \frac{N(N-1)}{2} p^{N-2} q^2$	$C_N^{N-1} p^{N-1} q = Np^{N-1} q$	$C_N^N p^N q^0 = p^N$

При составлении алгоритмов пословного машинного перевода и информационного поиска постоянно возникают задачи, связанные с прогнозированием появления в сегментах заданной длины определенного числа словоформ, морфем или словосочетаний, принадлежащих к некоторым классам. Формула Бернулли позволяет решать задачи этого типа, разумеется, при условии, что сохраняется принятое в п. 1 § 1 допущение о взаимной независимости образующих данный сегмент словоформ.

Рассмотрим в этой связи следующую задачу. Средняя длина простого предложения или синтаксически оформленной части сложного предложения в английских научно-технических текстах лежит между 10 и 11 словоформами. Относительная частота появления существительных в подъязыке английской электроники близка к 1/3 [6, с. 96—97]. Будем считать эту частоту априорной

вероятностью появления существительных в указанном подъязыке. Примем также, что типовым синтактически оформленным сегментом в английских научно-технических текстах является простое предложение, а также главное или придаточное предложение длиной в 10 словоформ.

Считая появление отдельных словоформ в этих сегментах независимыми событиями текста, определим вероятность того, что из 10 словоупотреблений, составляющих типовой сегмент, ровно два будут существительными.

Так как по условию $p = 1/3$, $q = 1 - p = 2/3$, $N = 10, x = 2$, то, пользуясь формулой (6.1), находим

$$P_N(x) = P_{10}(2) = C_{10}^2 \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^8 =$$

$$= \frac{10 \cdot 9}{1 \cdot 2} \cdot \frac{2^8}{3^{10}} = \frac{11520}{59049} = 0{,}1951 = 19{,}51\ \%.$$

Сохраняя те же условия и допущения, вычислим с помощью формулы Бернулли вероятности появления существительных в нашем типовом сегменте 0, 1, 2, ..., 10 раз. Результаты приведены в столбце (2) табл. 6.2.

Таблица 6.2

Вероятности появления существительных в английском предложении

x	$P_N(x)$	$P'_N(x)$	x	$P_N(x$	$P'_N(x)$
(1)	(2)	(3)	(1)	(2)	(3)
0	0,0173	0,0226	6	0,0569	0,0541
1	0,0867	0,0785	7	0,0163	0,0130
2	0,1951	0,1894	8	0,0031	0,0020
3	0,2601	0,2611	9	0,0003	0,0002
4	0,2276	0,2419	10	0,000020	0,000005
5	0,1366	0,1433			

Расчеты показывают, что в 10-словном предложении (сегменте) следует ожидать в среднем от двух до пяти форм имени существительного. Следовательно, на сегменты этого типа и должны быть ориентированы алгоритмы автоматического анализа английского текста. Появление сегментов с одним существительным или вообще без существительных, с одной стороны, а также с шестью, семью, восемью именными формами, с другой, маловероятно. И действительно, появление таких сегментов в английских научно-технических текстах хотя и возможно [ср. Comparing (10—13) to (10—4) it is seen that . . или Harvey Fletcher, Speech and Hearing in Communication, Bell Telephone Laboratories]*, но встречается

* Оба примера взяты из книги: H. Fletcher. Speech and Hearing in Communication. Toronto — New York—London, 1958, с. 1 и 172.

в виде исключения. Что же касается предложений, состоящих из девяти или десяти существительных, то такие отрывки просто невозможны. В рассмотренном теоретическом распределении вероятность появления таких сегментов практически равна нулю.

Часто, чтобы получить достаточно достоверные результаты, приходится производить большое число независимых испытаний. При этом величины N и x могут быть довольно велики, что делает вычисление по только что описанной схеме слишком трудоемким*. В таких случаях вычисление вероятностей $P_N(x)$ осуществляется по приближенным формулам, которые мы рассмотрим в § 3.

Иногда для решения лингвистической или информационной задачи необязательно определять все вероятности появления данного события 0, 1, 2, . . , N раз. Достаточно указать несколько наиболее вероятных или даже одно наивероятнейшее число появлений этого события.

Начнем с того, что опишем схему определения наивероятнейшего события Для этого рассмотрим поведение распределения (6.1). Из табл. 6.1 и 6.2 видно, что с увеличением x величина $P_N(x)$ возрастает и при некотором x_0 (оно называется *модальным значением*) достигает своего наибольшего значения $P_N(x_0)$. Затем по мере увеличения x вероятность $P_N(x)$ последовательно убывает. Чтобы определить модальное значение x_0, рассмотрим поведение функции $P_N(x)$ путем последовательного сравнения двух соседних членов распределения

Пусть $P_N(x_0)$ — наибольшее значение вероятности в распределении (6.1). Тогда должны выполняться следующие два неравенства:

$$P_N(x_0 - 1) \leqslant P_N(x_0),\ P_N(x_0) \geqslant P_N(x_0 + 1). \qquad (6.2)$$

Перепишем первое из неравенств (6.2) в виде

$$\frac{P_N(x_0)}{P_N(x_0-1)} = \frac{C_N^{x_0} p^{x_0} q^{N-x_0}}{C_N^{x_0-1} p^{x_0-1} q^{N-x_0+1}} = \frac{(N-x_0+1)p}{x_0 q} \geqslant 1. \qquad (6.3)$$

Заменив в последнем неравенстве q на $1 - p$, получаем

$$x_0 \leqslant Np + p. \qquad (6.4)$$

Аналогичным образом, записав второе из неравенств (6.2) в виде

$$\frac{P_N(x_0+1)}{P_N(x_0)} = \frac{C_N^{x_0+1} p^{x_0+1} q^{N-x_0-1}}{C_N^{x_0} p^{x_0} q^{N-x_0}} = \frac{(N-x_0)p}{(x_0+1)q} \leqslant 1, \qquad (6.5)$$

получим

$$x_0 \geqslant Np + p - 1. \qquad (6.6)$$

* Например, если бы мы захотели определить вероятность появления четырех существительных *напряжение* в серии из 2000 испытаний, зная, что в текстах по радиоэлектронике указанная словоформа согласно данным работы [6] имеет вероятность $p = 0{,}0023$, то эту вероятность мы должны были бы получить из равенства $P_{2000}(4) = C_{2000}^4 (0{,}0023)^4 (0{,}9977)^{1996}$, решение которого требует исключительно громоздких вычислений даже при условии использования специальных таблиц факториалов

Объединяя неравенства (6.4) и (6.6), приходим к двойному неравенству

$$Np + p - 1 \leqslant x_0 \leqslant Np + p. \qquad (6.7)$$

Левая часть неравенства (6.7) всегда на единицу меньше его правой части. Поэтому в тех случаях, когда $Np + p - 1$ и $Np + p$ являются дробными величинами, в качестве x_0 берется находящееся между ними целое число. Если же $Np + p - 1$ и $Np + p$ — целые числа, то x_0 имеет два целочисленных значения: $x_0 = Np + p - 1$ и $x_0' = Np + p$, которые выступают в качестве наивероятнейших значений появления данного события.

Теперь, пользуясь исходными данными об употребительности существительных, приведенными выше, определим наивероятнейшее число появлений существительных в английском 10-словном предложении.

Так как $N = 10$, $p = 1/3$, то, согласно (6.7), имеем

$$\frac{10}{3} - \frac{2}{3} < x_0 < \frac{10}{3} + \frac{1}{3}, \text{ или } 2\frac{2}{3} < x_0 < 3\frac{1}{3}.$$

Следовательно, наивероятнейшее число появлений существительных в 10-словном английском сегменте равно трем. Такой же результат дает распределение вероятностей, приведенное в табл. 6.2.

Зная модальное значение x_0, можно определить интересующее нас число вероятностей биномиального распределения. Вычисление их начинается обычно с определения максимальной вероятности $P_N(x_0)$:

$$P_N(x_0) = C_N^{x_0} p^{x_0} q^{N-x_0} = \frac{N!}{x_0!\,(N-x_0)!} p^{x_0} q^{N-x_0}. \qquad (6.8)$$

Вычисления остальных вероятностей производятся по следующим рекуррентным формулам, построенным на использовании выражений (6.3) и (6.5):

при $x < x_0$

$$\left.\begin{aligned}
P_N(x_0-1) &= \frac{x_0}{N-(x_0-1)} \frac{q}{p} P_N(x_0),\\
P_N(x_0-2) &= \frac{x_0-1}{N-(x_0-2)} \frac{q}{p} P_N(x_0-1),\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
P_N(x_{\min}+1) &= \frac{x_{\min}+2}{N-x_{\min}-1} \frac{q}{p} P_N(x_{\min}+2),\\
P_N(x_{\min}) &= \frac{x_{\min}+1}{N-x_{\min}} \frac{q}{p} P_N(x_{\min}+1),
\end{aligned}\right\} \qquad (6.9а)$$

при $x > x_0$

$$\left.\begin{aligned}
P_N(x_0+1) &= \frac{N-x_0}{x_0+1} \frac{p}{q} P_N(x_0),\\
P_N(x_0+2) &= \frac{N-(x_0+1)}{x_0+2} \frac{p}{q} P_N(x_0+1),\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
P_N(x_{\max}-1) &= \frac{N-(x_{\max}-2)}{x_{\max}-1} \frac{p}{q} P_N(x_{\max}-2),\\
P_N(x_{\max}) &= \frac{N-(x_{\max}-1)}{x_{\max}} \frac{p}{q} P_N(x_{\max}-1),
\end{aligned}\right\} \qquad (6.9б)$$

где $x_{\min} \geqslant 0$ и $x_{\max} \leqslant N$.

Только что описанный прием расчета биномиальных вероятностей можно проиллюстрировать следующим примером. В русском языке вероятность появления гласной в начале синтагмы или предложения составляет 23,21%. Пусть осуществлена повторная выборка в сто отдельных синтагм и предложений. Необходимо определить пять наивероятнейших частот появления начальной гласной, а также вычислить сумму их вероятностей.

Здесь $N = 100$, $p = 0{,}2321$, $q = 0{,}7679$; воспользовавшись неравенством (6.7), получаем

$$23{,}21 + 0{,}2321 - 1 < x_0 < 23{,}21 + 0{,}2321,$$

или

$$22{,}4421 < x_0 < 23{,}4421,$$

откуда $x_0 = 23$. Далее по формуле (6.8) вычисляем вероятность модального значения* x_0:

$$P_{100}(23) = \frac{100!}{23!\,77!} \cdot 0{,}2321^{23} \cdot 0{,}7679^{77} \approx 0{,}0943.$$

Затем по рекуррентным формулам (6.9) определяем значения $x_0 - 1$, $x_0 - 2$, $x_0 + 1$, $x_0 + 2$:

$$P_{100}(22) = \frac{23}{100-22} \cdot \frac{0{,}7679}{0{,}2321} \cdot 0{,}0943 = 0{,}0920,$$

$$P_{100}(21) = \frac{22}{100-21} \cdot \frac{0{,}7679}{0{,}2321} \cdot 0{,}092 = 0{,}0848,$$

$$P_{100}(24) = \frac{100-23}{23+1} \cdot \frac{0{,}2321}{0{,}7679} \cdot 0{,}0943 = 0{,}0914,$$

$$P_{100}(25) = \frac{100-24}{23+2} \cdot \frac{0{,}2321}{0{,}7679} \cdot 0{,}0914 = 0{,}0840.$$

* Решение этого примера осуществляется путем логарифмирования с использованием таблиц логарифмов факториала [9, с. 456].

Сумма полученных пяти вероятностей равна 0,45. Это значит, что при многократном извлечении из русских текстов серий синтагм и предложений, каждая из которых содержит 100 единиц, примерно половина из этих серий содержала бы от 21 до 25 предложений, начинающихся с гласного звука.

Только что описанный прием расчета вероятностей биномиального распределения имеет значительные преимущества перед последовательным вычислением, начиная с $x = 0$ (x_{min}). В последнем случае мы должны вычислять все значения $P_N(x)$, даже те, которые близки к нулю. При использовании только что описанной методики мы, получив максимальное значение $P_N(x)$, продолжаем вычисление вероятностей только до тех значений x_i и x_r ($x_i < x_0$, $x_r > x_0$), которые предусмотрены в условии задачи.

4. Полиномиальная схема. Если лингвистическое испытание имеет несколько исходов, то их вероятностное прогнозирование осуществляется с помощью полиномиальной схемы. Ее математическая модель строится следующим образом.

Предположим, что результатом некоторого лингвистического опыта может быть один из k различных попарно несовместимых исходов $A_1, A_2, \ldots, A_k$. Вероятность каждого из этих исходов обозначим соответственно через $P(A_1) = p_1$, $P(A_2) = p_2, \ldots, P(A_k) = p_k$. Так как событие $A_1 + A_2 + \ldots + A_k$ достоверно, то $p_1 + p_2 + \ldots + p_k = 1$. Осуществим N независимых испытаний и определим вероятности того, что событие A_1 появится x_1 раз, событие A_2 — x_2 раз, ..., событие A_k — x_k раз, где $x_1 + x_2 + \ldots + x_k = N$.

Указанный результат получается различными путями, каждый из которых соответствует различным перестановкам x_1 раз исхода A_1, x_2 раз исхода A_2, ..., x_k раз исхода A_k. Согласно теореме умножения вероятность появления каждой такой комбинации равна $p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$. Общее число этих комбинаций равно произведению $C_N^{x_1} C_N^{x_2} \ldots C_N^{x_k}$, которое приводится к выражению

$$\frac{N!}{x_1!\, x_2! \ldots x_k!}.$$

Отсюда получаем, что при N независимых испытаниях вероятность получить x_1 раз результат A_1, x_2 раз — результат A_2, ..., x_k раз — результат A_k равна

$$P_N(x_1, x_2, \ldots, x_k) = \frac{N!}{x_1!\, x_2! \ldots x_k!}\, p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}, \qquad (6.10)$$

где $0 \leqslant x_i \leqslant N$, а $\sum_{i=1}^{k} x_i = N$.

В том случае, когда $k = 2$, имеем

$$P_N(x_1, x_2) = \frac{N!}{x_1!\, x_2!}\, p_1^{x_1} p_2^{x_2}.$$

Учитывая, что $x_1 + x_2 = N$, а $p_1 + p_2 = 1$, и обозначая x_1 через x, x_2 — через $N - x$, p_1 — через p, а p_2 — через q, приходим к выражению

$$P_N(x) = \frac{N!}{x!\,(N-x)!}\, p^x q^{N-x} = C_N^x\, p^x q^{N-x},$$

т. е. к формуле Бернулли для простой схемы независимых испытаний. Формула Бернулли является, таким образом, частным случаем соотношения (6.10).

Используя только что описанную модель, определим вероятность того, что в 10-словном сегменте английского научно-технического текста появится ровно три существительных, две глагольных формы и пять словоформ, принадлежащих к другим классам (при этом мы снова пренебрегаем контекстными связями между словоформами, образующими рассматриваемый сегмент). Заданная нормой априорная вероятность появления существительных равна 0,33, вероятность глагольных форм составляет 0,16, а априорная вероятность остальных грамматических классов равна 0,51 [6, с. 104].

По условию задачи $N = 10$, $p_1 = 0{,}33$, $p_2 = 0{,}16$, $p_3 = 0{,}51$, $x_1 = 3$, $x_2 = 2$, $x_3 = 5$.

Применяя формулу (6.10), получаем

$$P_{10}(3,2,5) = \frac{10!}{3!\,2!\,5!} \cdot 0{,}33^3 \cdot 0{,}16^2 \cdot 0{,}51^5 \approx 0{,}0800.$$

Аналогичным образом можно рассчитать вероятность появления всех возможных количественных комбинаций существительных, глаголов и других классов слов в предложениях различной длины.

Как и простая схема, полиномиальная схема используется в повторных лингвистических выборках при условии, что величины N, $x_1, x_2, \ldots, x_k$ не слишком велики. При этих условиях использование рассмотренной схемы дает ценную информацию не только для вероятностного построения алгоритмов синтаксического анализа иностранного текста при машинном переводе. Эти алгоритмы позволяют также определять оптимальную последовательность подачи синтаксического материала при обучении иностранному языку в средней школе и вузе.

5. Пуассоновская схема. В лингвистической практике часто приходится иметь дело с такой речевой совокупностью, в которой составляющие ее тексты принадлежат к разным подъязыкам и стилям. Поскольку эти тексты строятся, исходя из различных норм, каждая лингвистическая единица имеет в каждом тексте свою априорную вероятность. В итоге вероятности появления и непоявления интересующих исследователя единиц меняются от опыта к опыту. Такая ситуация описывается *схемой Пуассона*. Математическая формализация этой схемы осуществляется в результате следующих рассуждений.

Пусть производится N независимых испытаний, в каждом из которых может появиться или не появиться событие A. Вероятности появления события A в 1, 2, ..., N испытаниях соответственно равны

$p_1, p_2, \ldots, p_N$, а вероятности его непоявления равны $q_1 = 1 - p_1$, $q_2 = 1 - p_2, \ldots, q_N = 1 - p_N$. Можно показать, что вероятность появления результата A в серии из N испытаний ровно x раз составляет

$$P_N(x) = p_1 p_2 p_3 \ldots p_x q_{x+1} \ldots q_N + \ldots + p_1 q_2 p_3 \ldots q_{N-1} p_N + \ldots$$
$$\ldots + q_1 q_2 q_3 \ldots q_{N-x} p_{N-x+1} p_{N-x+2} \ldots p_N. \qquad (6.11)$$

Таким образом, искомая вероятность представляет собой сумму всех возможных произведений, в каждом из которых p с разными индексами содержится x раз, а q с разными индексами входит $N-x$ раз.

Чтобы составить все возможные произведения из x вероятностей p_i и $N - x$ вероятностей q_i ($i = 1, 2, \ldots, N$), образуем произведения биномов

$$(q_1 + p_1 t)(q_2 + p_2 t) \ldots (q_N + p_N t) = \prod_{i=1}^{N} (q_i + p_i t), \qquad (6.12)$$

где t — некоторый произвольный параметр [10, с. 62].

Перемножая биномы и приводя подобные члены, приходим к равенству

$$\prod_{i=1}^{N} (q_i + p_i t) = \sum_{x=0}^{N} P_N(x) t^x,$$

в котором коэффициент при t^x есть не что иное, как выражение (6.11).

Раскрывая скобки в левой части равенства и приводя подобные члены, получим все вероятности $P_N(0)$, $P_N(1)$, $P_N(2)$, ..., $P_N(N)$, которые выступают в качестве коэффициентов соответственно при $t^0, t^1, t^2, \ldots, t^N$. Сумма всех вероятностей $P_N(x)$ равна единице:

$$\sum_{x=0}^{N} P_N(x) = 1.$$

В частном случае, когда $p_1 = p_2 = \ldots = p_N = p$, $q_1 = q_2 = \ldots = q_N = q$, имеем

$$(q + pt)^N = \sum_{x=0}^{N} C_N^x p^x q^{N-x} t^x,$$

откуда следует формула Бернулли.

Используя только что описанную математическую модель, решим следующую задачу.

Пусть производится повторная выборка именных групп из следующих четырех жанрово-тематических совокупностей русских текстов (подъязыков): записей непринужденной разговорной речи, поэзии, художественной прозы, научно-технических текстов. Именной группой считается словосочетание, в котором существительное стоит на последнем месте. Так, например, в предложении

используя только что описанную модель, решим следующую задачу триада *решим следующую задачу* будет именной группой.

Считая, что вероятность употребления существительных в разговорной речи равна 0,1, в поэзии — 0,2, в художественной прозе — 0,3, в научно-технических текстах — 0,4, найдем вероятности появления среди одновременно извлекаемых четырех словосочетаний ни одной, одной, двух, трех, четырех именных групп.

Событие, состоящее в появлении именной группы, по существу, соответствует событию, которое заключается в том, что из текста случайным образом выбирается форма имени существительного (к этой последней затем прибавляются два словоупотребления слева и тем самым формируется именная группа).

При четырех испытаниях, состоящих в извлечениях из каждого подъязыка по одному трехсловному сочетанию для нашего события, имеем вероятности: $p_1 = 0{,}1$, $p_2 = 0{,}2$, $p_3 = 0{,}3$, $p_4 = 0{,}4$. Для определения вероятностей $P_4(0)$, $P_4(1)$, $P_4(2)$, $P_4(3)$, $P_4(4)$ воспользуемся формулой (6.12). В результате получаем:

$$\prod_{i=0}^{4} (p_i t + q_i) = (0{,}1\,t + 0{,}9)(0{,}2\,t + 0{,}8)(0{,}3\,t + 0{,}7)(0{,}4t + 0{,}6) =$$
$$= 0{,}302 + 0{,}440\,t + 0{,}215\,t^2 + 0{,}040\,t^3 + 0{,}002\,t^4.$$

Из этого равенства следует, что вероятности получить в каждой серии 0, 1, 2, 3, 4 именных триады соответственно равны

$$P_4(0) = 0{,}302; \quad P_4(1) = 0{,}440; \quad P_4(2) = 0{,}215;$$
$$P_4(3) = 0{,}040; \quad P_4(4) = 0{,}002.$$

Схему Пуассона, как и две предыдущие схемы, целесообразно применять к лингвистическим испытаниям тогда, когда мы можем организовать повторную выборку, а величины N и x не очень велики.

В предыдущих разделах мы научились прогнозировать исходы массовых лингвистических испытаний. Такие прогнозы мы можем пока осуществлять применительно к повторным выборкам, опираясь на классическое определение вероятности, т. е. при условии, что опыт осуществляется относительно сравнительно ограниченной по объему конечной совокупности лингвистических объектов. Такая ситуация встречается в лингвистике сравнительно редко. Чаще всего языковеду приходится иметь дело с бесповторной выборкой, исследующей редко встречающиеся лингвистические единицы. В этих условиях распределение вероятностей появления события A подчиняется гипергеометрическому закону [6, с. 156—162].

6. Бесповторная лингвистическая выборка и ее описание с помощью формулы Бернулли. Гипергеометрический закон может применяться только к конечным генеральным совокупностям, объем которых известен. Поскольку в лингвистических задачах объем гене-

ральной совокупности текстов, порождаемых открытой системой языка, обычно не является конечной величиной, применение указанного закона для прогнозирования исхода лингвистических опытов в бесповторных выборках оказывается нереальным. Вместе с тем при определенных условиях гипергеометрическая вероятность хорошо аппроксимируется биномиальной вероятностью [30, кн. 1, с. 271 и сл.]. Поэтому, не боясь нарушения математической строгости, мы будем производить расчет вероятностей появления события A ровно x раз в нашей бесповторной выборке так, как если бы речь шла о повторной выборке. Иными словами, мы применяем к бесповторным выборкам биномиальный закон.

Будем рассматривать данные S текстов как S серий или выборок, каждая из которых состоит из N независимых испытаний. Лингвистическое событие A может появиться в каждой серии x раз ($x = 0, 1, 2, \ldots, N$). Нетрудно заметить, что имеются группы серий, в которых A появляется 0, 1, 2, ..., N раз. Отсюда следует, что частость появления события A ровно x раз в одной серии определяется отношением $f_N(x) = S_x/S$, где S_x — количество серий, в которых событие A появится ровно x раз.

Априорная вероятность появления события A в одной наугад взятой серии равна

$$p \approx \frac{\Sigma x\, S_x}{NS}$$

и, следовательно,

$$q \approx 1 - \frac{\Sigma x\, S_x}{NS}.$$

В получаемом теоретическом распределении каждому значению x соотнесена не его вероятность, а некоторое теоретически ожидаемое число серий (выборок) S_x^T, в которых событие A появляется ровно x раз. Поскольку

$$S_x^T = SP_N(x) = SC_N^x p^x q^{N-x}, \qquad (6.13)$$

нетрудно заметить, что величины S_x^T и $P_N(x)$ связаны коэффициентом пропорциональности S.

Для иллюстрации вышесказанного рассмотрим следующий пример.

При определении стилистико-статистических особенностей употребления существительных в романе М. Ауэзова «Путь Абая» (на казахском языке) из текста романа случайным образом выбирается 400 отрезков текста по 25 словоупотреблений каждый. Данные о частотах употребления существительных в этих отрезках показаны в столбцах (1) и (2) табл. 6.3. Необходимо вычислить теоретическое биномиальное распределение вероятностей появления x существительных в одной серии.

Таблица 6.3

Вероятностно-статистические характеристики употребления существительных в романе М. Ауэзова «Путь Абая»

Число появлений события x	Эмпирические частоты появления выборок S_x	Теоретически ожидаемое число выборок S_x^T	S_x^T, округленное до целых чисел	Частость $f_N(x) = \frac{S_x}{S}$	Вероятность появления события ровно x раз
(1)	(2)	(3)	(4)	(5)	(6)
0	—	0,050	1	—	0,0001
1	3	0,580		0,0075	0,0014
2	5	2,960	3	0,0125	0,0074
3	7	9,710	10	0,0175	0,0243
4	24	22,880	23	0,0600	0,0572
5	40	41,210	41	0,1000	0,1030
6	52	58,870	59	0,1300	0,1472
7	64	68,480	68	0,1600	0,1712
8	66	66,030	66	0,1650	0,1651
9	47	53,440	53	0,1175	0,1336
10	48	36,640	37	0,1200	0,0916
11	24	21,440	21	0,0600	0,0536
12	14	10,720	11	0,0350	0,0268
13	3	4,600	5	0,0075	0,0115
14	2	1,680	2	0,0050	0,0042
15	1	0,520		0,0025	0,0013
16	—	0,140		—	0,0004
17	—	0,030		—	0,0001
	$\Sigma S_x = 400$	399,980	400	1,0000	1,0000

Здесь $S = 400$, $N = 25$. Используя произведения величин x и S_x, приведенных в первых двух столбцах табл. 6.3, находим

$$p = \frac{\Sigma x\, S_x}{NS} = \frac{1}{N}\left[\frac{\Sigma x\, S_x}{S}\right] = \frac{7{,}59}{25} = 0{,}3038.$$

Приняв $p \approx 0{,}3$ и $q \approx 0{,}7$, согласно (6.7), имеем

$$25 \cdot 0{,}3 - 0{,}7 < x_0 < 25 \cdot 0{,}3 + 0{,}3, \quad \text{или} \quad 6{,}8 < x_0 < 7{,}8,$$

откуда следует, что $x_0 = 7$. Тогда

$$P_N(x_0) = P_{25}(7) = C_{25}^7 0{,}3^7 \cdot 0{,}7^{18}.$$

Логарифмированием находим, что $P_{25}(7) = 0{,}17119$. Следовательно,

$$S_x^T = SP_{25}(7) = 400 \cdot 0{,}17119 \approx 68{,}480.$$

Остальные значения ожидаемого числа выборок, вычисленные с помощью соотношений (6.9) и (6.13), приведены в табл. 6.3.

7. **Определение вероятности появления лингвистического события от a до b раз.** Определение вероятности того, что та или иная лингвистическая единица появится в данной выборке ровно x раз, пред-

ставляет обычно небольшой интерес с точки зрения языкознания. Гораздо важнее уметь вычислять вероятность появления лингвистического события от a до b раз в заданном массиве текста.

Пусть B'_x — событие, состоящее в том, что лингвистическая единица A появится не менее a и не более b раз. Тогда вероятность $P_N\ (a \leqslant x \leqslant b)$ этого события составляет

$$P_N\ (a \leqslant x \leqslant b) = P_N\ (a) + P_N\ (a+1) + \ldots + P_N\ (b-1) + P_N\ (b) =$$

$$= \sum_{x=a}^{b} P_N\ (x) = \sum_{x=a}^{b} C_N^x\, p^x\, q^{N-x}.$$

Если количество членов, отвечающих значениям x от a до b, намного больше общего количества членов, соответствующих значениям x от 0 до $a-1$ и от $b+1$ до N, то удобнее проводить суммирование вероятностей по этим двум последовательностям, получая тем самым вероятность события $\overline{B}'_x$, противоположного событию B'_x:

$$P\ (\overline{B}'_x) = \sum_{x=0}^{a-1} C_N^x\, p^x\, q^{N-x} + \sum_{x=b+1}^{N} C_N^x\, p^x\, q^{N-x}.$$

Искомая же вероятность равна

$$P_N\ (a \leqslant x \leqslant b) = 1 - P\ (\overline{B}'_x) =$$

$$= 1 - \sum_{x=0}^{a-1} C_N^x\, p^x\, q^{N-x} - \sum_{x=b+1}^{N} C_N^x\, p^x\, q^{N-x}. \qquad (6.14)$$

Рассмотрим некоторые частные случаи. Предположим, что необходимо определить вероятность того, что лингвистическая единица A встретится не менее a раз. Здесь

$$P_N\ (x \geqslant a) = \sum_{x=a}^{N} C_N^x\, p^x\, q^{N-x}.$$

Если a мало, то целесообразно пользоваться выражением

$$P_N\ (x \geqslant a) = 1 - \sum_{x=0}^{a-1} C_N^x\, p^x\, q^{N-x},$$

являющимся частным случаем формулы (6.14).

В том случае, когда $a = 1$, имеем

$$P_N\ (1 \leqslant x \leqslant N) = 1 - C_N^0\, p^0\, q^N = 1 - q^N. \qquad (6.15)$$

Пусть, например, из русских текстов по радиоэлектронике случайным образом извлечено 1000 словоупотреблений. Найдем вероятность того, что словоформа *напряжение* встретится хотя бы один раз, если ее частость, согласно данным работы [6, с. 175—176], равна 0,0023.

Здесь $N = 1000$, $p = 0{,}0023$, $q = 0{,}9977$. Пользуясь формулой (6.15), находим

$$P_N(x \geqslant 1) = P_{1000}\ (1 \leqslant x \leqslant 1000) = 1 - 0{,}9977^{1000} \approx$$
$$\approx 1 - 0{,}10 = 0{,}90.$$

Это означает, что если осуществить 100 выборок по 1000 словоупотреблений каждая, то появление словоформы *напряжение* можно наверняка ожидать в 90 выборках.

Вероятность появления события A не более b раз также определяется путем суммирования вероятностей, в которых событие появится 0, 1, 2, ..., b раз:

$$P_N\ (x \leqslant b) = \sum_{x=0}^{b} C_N^x\, p^x\, q^{N-x}.$$

При близком к N значении b эту вероятность следует вычислять с помощью формулы

$$P_N\ (x \leqslant b) = 1 - \sum_{x=b+1}^{N} C_N^x\, p^x\, q^{N-x}, \qquad (6.16)$$

также представляющей собой частный случай выражения (6.14).

Чтобы проиллюстрировать описанную методику, определим вероятность того, что в извлеченном наугад из романа М. Ауэзова «Путь Абая» отрывке в 25 словоупотреблений будет не более шестнадцати существительных.

Здесь $N = 25$, $p = 0{,}3$, $q = 0{,}7$ (см. п. 6); вместо того чтобы определять, а затем суммировать вероятности появления 0, 1, 2, ..., 16 существительных, определим вероятность появления семнадцати существительных (появление более чем семнадцати существительных практически равно нулю — см. табл. 6.3):

$$P_{25}\ (17) = C_{25}^{17} \cdot 0{,}3^{17} \cdot 0{,}7^8 = 0{,}0004.$$

Тогда искомая величина, согласно формуле (6.16), равна

$$P\ (x \leqslant 16) = 1 - P_{25}\ (17) = 1 - 0{,}0004 = 0{,}9996.$$

Иными словами, если взять 10000 выборок по 25 словоупотреблений, то в 9996 выборках можно ожидать появление не более 16 существительных.

8. Определение необходимого объема выборки. В лингвистических исследованиях и особенно при подготовке лингвистических программ машинного перевода и информационного поиска постоянно возникает потребность определять объем выборки, необходимый для того, чтобы обеспечить с заданной вероятностью появление хотя бы один раз интересующей нас лингвистической единицы.

Для этого приведем сначала соотношение

$$P_N\ (1 \leqslant x \leqslant N) = 1 - q^N = 1 - (1-p)^N$$

к виду

$$(1-p)^N = 1 - P_N\ (1 \leqslant x \leqslant N).$$

Прологарифмировав обе части равенства и произведя необходимые преобразования, получим

$$N = \frac{\lg [1 - P_N (1 \leqslant x \leqslant N)]}{\lg (1-p)}, \qquad (6.17)$$

где N указывает на необходимый объем выборки.

Пусть, например, нужно определить тот объем выборки русских текстов по радиоэлектронике, который необходим для того, чтобы с вероятностью в 90% словоформа *напряжение* появилась в нем хотя бы один раз.

Здесь $p = 0{,}0023$ (см. п. 7), $P_N\ (1 \leqslant x \leqslant N) = 0{,}90$. По формуле (6.17) находим

$$N = \frac{\lg (1-0{,}90)}{\lg (1-0{,}0023)} = \frac{\lg 0{,}10}{\lg 0{,}9977} = \frac{-1}{-0{,}001} = 1000.$$

Иными словами, для того, чтобы с уверенностью в 90% утверждать, что словоформа *напряжение* встретится хотя бы один раз, нужно просмотреть выборку длиной в тысячу словоупотреблений.

§ 2. Случайная лингвистическая величина, ее характеристики и функция распределения

1. Дискретные и непрерывные случайные величины в речевой деятельности. При проведении лингвистического опыта мы постоянно встречаемся с такими величинами, численные значения которых невозможно раз навсегда определить, причем эти значения меняются под влиянием случайных воздействий. Так, например, выбирая наугад слова из текста, мы встречаем слова длиной в 1, 2, 3 и т. д. букв. Эти слова могут содержать 0, 1, 2, 3 и т. д. гласных или согласных фонем. Длина слова, количество гласных или согласных фонем выступают в качестве *случайных* величин, т.е. таких лингвистических величин, которые могут в результате испытаний принимать различные, заранее непредсказуемые значения.

Только что рассмотренные случайные лингвистические величины принимают определенные четко отграниченные друг от друга прерывные значения. Такие величины называются *дискретными случайными величинами*.

Если же случайная величина принимает сплошь все значения в каком-то интервале на числовой оси, то мы имеем дело с *непрерывной случайной величиной*. Примером непрерывной случайной величины является интенсивность звука, которая колеблется обычно в определенных пределах (ср. ниже п. 5).

Когда фонолог, грамматист или лексиколог исследует структуру планов содержания и выражения, он всегда имеет дело с дискретными случайными величинами. Однако, обращаясь к фонетическим и семантическим исследованиям — исследованиям, касающимся субстанции планов выражения и содержания, лингвист должен оперировать непрерывными случайными величинами.

Принятие случайной величиной X конкретного значения x_i есть случайное событие. Поэтому описываемые ниже свойства случайной лингвистической величины являются в какой-то степени обобщением того, что было сказано о случайном лингвистическом событии.

2. Законы распределения дискретной случайной величины. Чтобы полностью задать случайную величину, недостаточно только указать те значения, которые она может принимать. Необходимо еще знать для каждого значения x_i ту вероятность $P\ (X = x_i) = p_i$, с которой случайная величина принимает это значение. Если случайная величина X является дискретной и принимает возможные значения $x_0, x_1, x_2, \ldots, x_N$, то вероятности p_i точно соответствуют вероятностям $P_N\ (x)$, с которыми мы имели дело в § 1, п. 3.

Рассматриваемые нами возможные значения случайной величины являются событиями попарно несовместимыми и образуют полную группу событий, поэтому сумма их вероятностей равна единице, т. е. $\Sigma p_i = 1$.

Правило, связывающее значения случайной величины и соответствующие им вероятности, носит название з а к о н а р а с п р е д е л е н и я дискретной случайной величины.

Простейшей формой закона является т а б л и ц а р а с п р е д е л е н и я или, как еще называют, р я д р а с п р е д е л е н и я. В этой таблице перечисляются все возможные значения случайной величины и указываются соответствующие им значения вероятностей; такова, например, табл. 6.1, в которой представлено биномиальное распределение. Закон распределения может быть задан в виде формулы: примером аналитического выражения биномиального распределения служит формула (6.1). Наконец, для передачи закона распределения можно использовать графическую иллюстрацию; соответствующие примеры будут приведены ниже.

3. Понятие функции распределения случайной величины. Поскольку языкознание имеет дело не только с дискретными, но и с непрерывными величинами, необходимо наряду с распределениями дискретных величин рассмотреть также распределения непрерывных случайных величин. Это тем более необходимо потому, что нам придется представлять распределения некоторых дискретных величин в виде непрерывных распределений (см. ниже § 3, п. 4).

Если дискретную случайную величину характеризует таблица, в которой указываются все значения этой величины и ее вероятности, то для непрерывной случайной величины такую таблицу построить нельзя: во-первых, непрерывная случайная величина принимает бесконечное множество значений; во-вторых, вероятность того, что рассматриваемая непрерывная случайная величина точно примет то или иное численное значение, равна нулю (см. ниже, п. 5).

В связи с этим встает вопрос об отыскании такой модели распределения, которая характеризовала бы как дискретную, так и непрерывную случайную величину.

Строя такую модель, воспользуемся неравенством $X < x$, где x является переменной величиной. Это неравенство означает,

что случайная величина X принимает всевозможные значения, меньшие чем x. Вероятность появления такой величины равна $P\ (X < x)$.

Ясно, что вероятность принимаемых значений случайной величины зависит от значений переменной x. Поэтому указанная вероятность является некоторой функцией от x. Обозначив эту функцию как

$$F\ (x) = P\ (X < x),$$

или иначе,

$$F\ (x) = P\ (-\infty < X < x),$$

будем называть ее *функцией распределения*, или *интегральной функцией распределения случайной величины*. Ее называют также *кумулятивной функцией*, т. е. такой функцией, значения которой представляют собой каждый раз сумму численности данного признака и численностей всех предшествующих ему признаков.

Функция распределения случайной величины обладает следующими свойствами:

1. *Так как значение вероятности заключено между* 0 *и* 1, *то справедливо неравенство* $0 \leqslant F\ (x) \leqslant 1$.

2. *Если случайная величина ограничена, т. е. принимает все возможные значения в некотором отрезке* $[a, b]$, *то для всех значений* X, *меньших чем* a (*невозможное событие*) $F\ (x) = 0$, *а для всех значений, больших чем* b (*достоверное событие*) $F\ (x) = 1$.

3. *Если случайная величина принимает любые значения в промежутке* $-\infty < X < +\infty$, *то имеют место равенства:*

$$\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to +\infty} F(x) = 1.$$

4. *Вероятность того, что случайная величина* X *примет значение, удовлетворяющее неравенству* $x_1 \leqslant X < x_2$, *равна приращению ее функции распределения* $F\ (x)$ *на интервале от* x_1 *до* x_2, *т. е.*

$$P\ (x_1 \leqslant X < x_2) = F\ (x_2) - F\ (x_1). \qquad (6.18)$$

5. *Функция распределения случайной величины является неотрицательной неубывающей функцией аргумента, т. е. при* $x_1 < x_2$ *имеет место неравенство* $F\ (x_1) \leqslant F\ (x_2)$.

Все перечисленные свойства характеризуют функцию распределения как для дискретных, так и для непрерывных случайных величин.

4. Функция распределения для случайной лингвистической величины дискретного типа. Используя данные о теоретическом распределении частот существительных в английской научно-технической речи (см. § 1, п. 3), построим функцию распределения $F\ (x)$.

По условию случайная величина $X < x$ принимает все целочисленные значения, заключенные на отрезке [0, 10]. Исходя из свойства 2, а также учитывая свойства 1, 4, 5, определяем кумулятивную функцию $F\ (x)$ как сумму вероятностей случайной величины X, не превосходящей x:

$$F\ (x) = P\ (X < 1) = P\ (X = 0) = 0{,}01734;$$

$$F\ (x) = P\ (X < 2) = P\ (X = 0) + P\ (X = 1) = 0{,}01734 + 0{,}08671 = 0{,}10405.$$

Аналогичным образом имеем:

$$F\ (x) = P\ (X < 3) = 0{,}29914; \qquad F\ (x) = P\ (X < 4) = 0{,}55926;$$
$$F\ (x) = P\ (X < 5) = 0{,}78686; \qquad F\ (x) = P\ (X < 6) = 0{,}92342;$$
$$F\ (x) = P\ (X < 7) = 0{,}98032; \qquad F\ (x) = P\ (X < 8) = 0{,}99658;$$
$$F\ (x) = P\ (X < 9) = 0{,}99963; \qquad F\ (x) = P\ (X < 10) = 0{,}99965;$$

если $x > 10$, то $F\ (x) = P\ (X < x) = 1{,}0000$.

График рассматриваемой функции показан на рис. 37. В том случае, когда случайная величина принимает значения, меньшие чем $x = 0$, функция $F\ (x)$ равна нулю. Этому значению соответствует линия, лежащая на оси абсцисс левее начала координат.

Как только случайная величина X примет значение, равное нулю (т. е. при условии, что $0 < x \leqslant 1$ и $X < x$), функция $F\ (x) = P\ (X < x)$ получает значение 0,0173. Иными словами, в точке A_0 эта функция претерпевает разрыв, сопровождающийся скачком ее численного значения. Величина этого скачка в точности равна значению вероятности $P_0 = A_0B_0 = 0{,}0173$. Это значение функция $F\ (x) = P\ (X < x)$ сохраняет вплоть до следующего целочисленного значения случайной величины, т. е. до точки A_1, где $F\ (x) = P\ (X < 2)$. Здесь снова происходит скачок величиной в $P_1 = A_1B_1 = 0{,}0867$, причем $F\ (x) = P\ (X < 2) = P_0 + P_1 = 0{,}1040$. Это значение $F\ (x)$ сохраняется в полуоткрытом промежутке [1, 2), после чего происходит новый скачок $A_2B_2 = P_2$ и т. д. В силу того что при $x \geqslant 10$ значение кумулятивной функции равно единице, график ее при $x \geqslant 10$ сливается с прямой $F\ (x) = 1$.

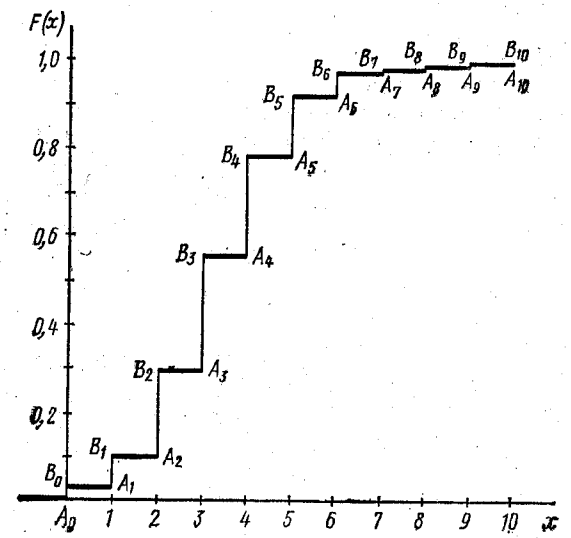


Рис. 37

Приведенный пример показывает, что функция распределения любой дискретной случайной величины всегда является разрывной ступенчатой функцией, скачки которой происходят в точках, соответствующих значениям случайной величины X. Скачок в каждой такой точке равен вероятности P_i того, что случайная величина примет целочисленное значение i. Сумма всех скачков равна единице.

5. Функция распределения для случайной лингвистической величины непрерывного типа. Прежде чем говорить об особенностях описания непрерывной случайной величины, рассмотрим следующий пример. Пусть произведен статистический эксперимент, ставивший целью измерения чувствительности слуха относительно звукового тона в 1500 Гц [6, с. 122—123]. Испытанию было подвергнуто две тысячи испытуемых. Измерения велись путем постепенного повышения уровня интенсивности сигнала, начиная от «звуков», не воспринимаемых ухом человека, к слышимым звукам. На каждый распознанный сигнал испытуемый отвечал включением светового сигнала. Таким образом определялся тот уровень интенсивности (в децибелах, сокращенно — дБ), при котором каждый испытуемый начинает слышать звук указанного тона. Само собой разумеется, что разные испытуемые начинают слышать звук на разных уровнях его интенсивности. Такие разные уровни интенсивности, необходимые для того, чтобы тот или иной взрослый человек услышал звук заданного тона, могут рассматриваться как значения случайной величины X. Закон распределения этой случайной величины по интервалам, ширина которых определяется условиями эксперимента, показан в табл. 6.4.

Таблица 6.4

X(дБ)	$x<-13,5$	$-13,5\leqslant X<-10,5$	$-10,5\leqslant X<-7,5$	$-7,5\leqslant X<-4,5$	$-4,5<X<-1,5$	$-1,5\leqslant X<1,5$
P_i	0,00	0,01	0,02	0,08	0,25	0,31
X (дБ)	$1,5\leqslant X<4,5$	$4,5\leqslant X<7,5$	$7,5\leqslant X<10,5$	$10,5\leqslant X<13,5$	$X\geqslant 13,5$	
P_i	0,29	0,08	0,02	0,01	0,00	

Только что построенный закон распределения существенно отличается от закона распределения случайной величины дискретного типа. Действительно, если последняя принимала конечное или во всяком случае счетное множество значений, то теперь наша случайная величина — уровень интенсивности — может принимать бесконечное множество значений.

Один испытуемый может воспринять звук с интенсивностью в —13,5 дБ, другой услышит звук при условии, что его сила равна —13,49 дБ, для третьего испытуемого этот уровень будет равен —13,489 дБ и т. д.

Указать в таблице распределения все бесконечное множество значений случайной величины невозможно. Поэтому и приходится говорить об интервалах, в которые могут попадать ее значения. При этом отмечаются либо границы интервала, либо указывается его середина (начало, конец). Вероятности, приписываемые каждому из интервалов, являются вероятностями того, что рассматриваемая непрерывная величина попадет в данный интервал.

При графическом изображении интервального ряда распределения значений случайной лингвистической величины целесообразно пользоваться *гистограммой*, представляющей собой последовательность прямоугольников, основание которых равно ширине интервала, а высота — соответствующей этому интервалу вероятности (рис. 38). Нетрудно заметить, что прямоугольники гистограммы образуют фигуру, ограниченную сверху ломаной линией, а снизу — прямой KL. Площадь этой фигуры, представляющая сумму площадей прямоугольников, равна единице.

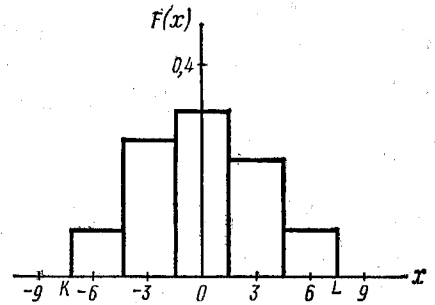


Рис. 38

Интегральный ряд распределения непрерывной случайной величины может быть представлен также в виде ступенчатого кумулятивного графика, аналогичного тому графику, который мы строили для кумулятивной функции $F(x)$ дискретной случайной величины (см. рис. 37). Кумулятивный график закона распределения уровней интенсивности звукового тона показан на рис. 39.

С теоретической точки зрения интервальное представление непрерывной случайной величины не дает достаточно адекватного ее описания. Сам выбор ширины границ интервала всегда произволен, поведение случайной величины внутри этого интервала остается неопределенным. Наконец, непрерывность рассматриваемой величины не находит отражения в дискретном характере интервального ряда и соответствующих ему графиков.

Чтобы избежать этих затруднений, необходимо использовать особый математический аппарат. Прежде чем вводить этот аппарат, рассмотрим в геометрической интерпретации поведение непрерывной случайной величины. Как уже говорилось, ширина интервалов в распределении непрерывной случайной величины выбирается произвольно. Теоретически ничто не мешает нам последовательно уменьшать эти интервалы, как это мы делали в гл. 2, § 2. При этом отрезки ломаной линии, ограничивающей сверху фигуру, изображенную на рис. 38, становятся все меньше, пока ломаная линия

не превратится в плавную кривую (рис. 40), которую называют *дифференциальной кривой распределения*. При этом сумма площадей прямоугольников, равная единице, практически не будет уже от-

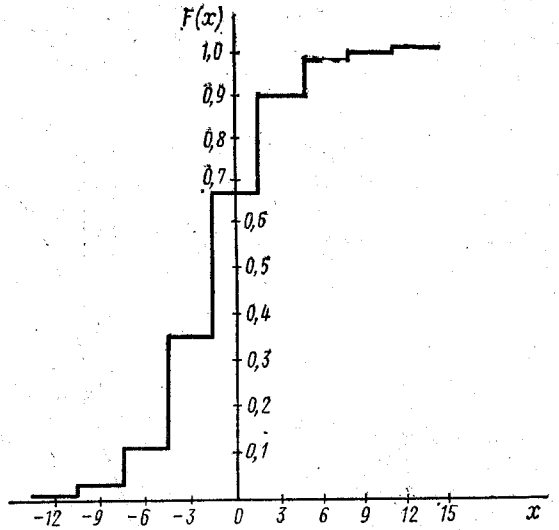


Рис. 39

личаться от площади фигуры, ограниченной снизу отрезком KL, а сверху — кривой, полученной из ломаной линии.

Аналогичная картина наблюдается и в кумулятивном графике. По мере уменьшения ширины интервала будет расти число интер-

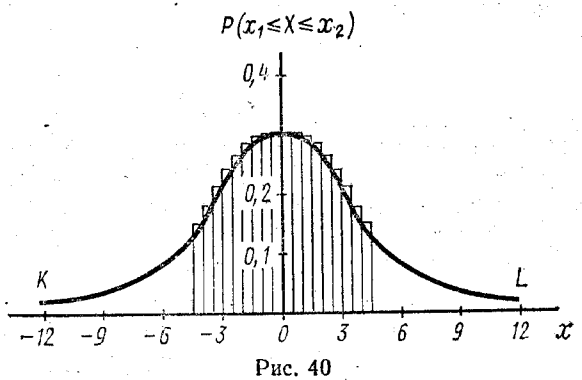


Рис. 40

валов и скачков между ними с одновременным уменьшением величины этих скачков до тех пор, пока ступенчатая линия не превратится в плавную кривую линию, которую принято называть *интегральной кривой распределения* (рис. 41).

Теперь займемся аналитическим представлением поведения непрерывной случайной величины.

Ранее (см. табл. 6.4) мы говорили о том, что вероятности, приписываемые интервалам, указывают на ту вероятность, с которой случайная величина X попадает в заданный интервал. Рассматривая каждый интервал как полуоткрытый промежуток $[x_1, x_2)$, мы можем считать, что попадание в него случайной величины X равносильно выполнению неравенства $x_1 \leqslant X < x_2$. Согласно свойству 4 интегральной функции распределения вероятность выполнения этого неравенства равна

$$P(x_1 \leqslant X < x_2) = F(x_2) - F(x_1).$$

Иными словами, *вероятность попадания случайной величины в заданный интервал равна приращению функции на этом интервале.*

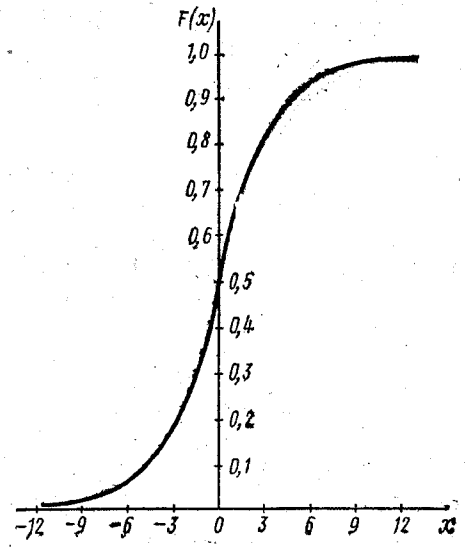


Рис. 41

Если неограниченно уменьшать интервал $[x_1, x_2)$, как это мы уже делали при построении кривой распределения, то вместо вероятности того, что случайная величина попадает на этот участок, мы получаем в пределе, что величина X примет отдельно взятое значение x_1:

$$P(X = x_1) = \lim_{x_2 \to x_1} P(x_1 \leqslant X < x_2) = \lim_{x_2 \to x_1} [F(x_2) - F(x_1)]. \quad (6.19)$$

Поскольку функция $F(x)$ не имеет разрывов и непрерывна во всех точках, в том числе и в точке x_1, предел (6.19) равен нулю*.

Из всего сказанного следует, что *вероятность каждого отдельного значения непрерывной случайной величины равна нулю*, т. е.

$$P(X = x) = 0.$$

* При решении практических задач попадание непрерывной случайной величины X в отдельную точку реального смысла не имеет. Ведь абсолютно точное значение физической величины — в нашем случае длины, высоты, интенсивности звука человеческой речи — является лишь математической абстракцией. На практике в результате измерений мы получаем интервалы, равные той наименьшей единице, которую может показать измерительный прибор. Поэтому наблюдаемые в опыте значения случайной величины, строго говоря, всегда дискретны. Но зная, что по своей внутренней природе рассматриваемая случайная величина непрерывна, мы применяем для ее описания непрерывное распределение.

Только что сформулированное свойство функции распределения может показаться лингвисту парадоксальным. С одной стороны, приводя классическое определение вероятности, мы говорили, что события, имеющие нулевую вероятность, — это невозможные события; с другой стороны, из всего только что сказанного вытекает, что событие, состоящее в том, что случайная величина X примет значение x, хотя и возможно, но имеет вероятность, равную нулю.

Между тем появление событий, обладающих нулевой вероятностью, можно представить себе при двух условиях: во-первых, эти события должны рассматриваться вне рамок классической схемы конечного числа случаев, во-вторых, они должны изучаться по статистической схеме или в рамках аксиоматического построения теории вероятностей (см. гл. 5, § 3).

6. Семантическая интерпретация непрерывной случайной величины. Некоторые области субстанции содержания представляют собой непрерывный континуум [52, с. 308—318]. Это значит, что между некоторыми родственными «универсальными» значениями нет четких границ, и между ними всегда можно найти бесконечное число переходных смысловых оттенков. Особенно наглядно эта ситуация прослеживается в непрерывности цветового спектра. Хотя разные языки по-своему формируют алфавиты (парадигмы) слов, обозначающие отдельные хроматические интервалы [44], в каждом языке можно найти средства для обозначения все более тонких оттенков цветов, т. е. последовательно сужать эти семантические интервалы*. Так, например, внутри зеленого можно выделить синевато-зеленый оттенок, внутри синевато-зеленого можно найти синевато-зеленый с серым оттенком цвет, затем можно выделить синевато-зеленый с водянисто-серым оттенком цвет и т. д.

Теперь поставим опыт, состоящий в том, что из книги, в которой дается описание разных оттенков цвета, наугад выбираются предложения и определяется, о каком цвете или оттенке в них идет речь. Если говорить о таких занимающих широкие хроматические интервалы цветах, как зеленый, то вероятность того, что они окажутся упомянутыми в наугад взятом предложении, является достаточно большой. Однако по мере раздробления нашего цветового спектра на все более частные оттенки вероятность появления обозначений каждого из них будет уменьшаться. В конце концов мы придем к тому, что хотя в каждом наугад взятом предложении и будет говориться о каком-то цветовом оттенке, но вероятность появления конкретного оттенка будет равна нулю.

Таким образом, если считать непрерывной случайной величиной X некоторое цветовое значение, то окажется, что при осуществлении нашего опыта непрерывная случайная величина обязательно примет одно из своих возможных значений, хотя до опыта вероятность появления каждого из них была равна нулю. Иными словами,

* В одном из специальных английских словарей приводится около четырех тысяч названий оттенков цвета [6, с.130], число этих названий можно увеличивать и дальше.

осуществится одно из событий, вероятность появления которого равна нулю. Поскольку частота события не равна, а лишь приближается в большом количестве опытов к вероятности, то утверждение согласно которому вероятность события $X = x$ равна нулю, означает лишь, что при многократном повторении опыта это событие будет осуществляться сколь угодно редко.

7. Плотность распределения вероятностей. Применительно к дискретным случайным величинам функция распределения является такой функцией, с помощью которой суммируются значения P_i, выступающие в качестве элементов вероятности. Выясним теперь, что является элементом вероятности непрерывных случайных величин.

Для этого рассмотрим интервал $[x, x+\Delta x]$ и определим вероятность того, что случайная величина X попадет в этот интервал. Согласно свойству 4 функции распределения имеем

$$P\ (x \leqslant X < x + \Delta x) = F\ (x + \Delta x) - F\ (x).$$

Разделив вероятность $P\ (x \leqslant X < x + \Delta x)$ на длину интервала Δx, получаем величину вероятности, приходящуюся на единицу длины этого интервала:

$$\frac{1}{\Delta x} P\ (x \leqslant X < x + \Delta x).$$

Эту величину будем называть *средней плотностью вероятности* на данном интервале.

Если последовательно уменьшать интервал Δx, то в пределе получим функцию

$$f(x) = \lim_{\Delta x \to 0} \frac{P\ (x \leqslant X < x + \Delta x)}{\Delta x}, \qquad (6.20)$$

называемую *плотностью вероятности*, или *плотностью распределения вероятностей*.

Если в правую часть равенства (6.20) вместо числителя подставить приращение функции $F\ (x + \Delta x) - F\ (x)$, то получим выражение

$$f(x) = \lim_{\Delta x \to 0} \frac{F\ (x+\Delta x) - F\ (x)}{\Delta x} = F'\ (x),$$

из которого следует, что плотность вероятности является производной от функции распределения (см. гл. 3, § 1).

Отсюда следует, что функцию распределения можно определить и через плотность вероятности. С одной стороны, по формуле (4.25) имеем

$$\int_a^b f(x)\, dx = F(b) - F(a).$$

С другой стороны, в силу свойства 4 функции распределения находим

$$P(a \leqslant X \leqslant b) = F(b) - F(a) = \int_a^b f(x)\,dx.$$

Полагая $a = -\infty$, $b = x$, приходим к функции распределения:

$$F(x) = P(-\infty < X < x) = \int_{-\infty}^{x} f(x)\,dx. \qquad (6.21)$$

Таким образом, дифференциальная функция $f(x)$ и интегральная функция $F(x)$ взаимно определяют друг друга.

Плотность распределения, так же как и функция распределения, является одной из форм закона распределения. Однако, если функция распределения является универсальной характеристикой закона распределения как дискретных, так и непрерывных случайных величин, то плотность распределения характеризует только непрерывные случайные величины.

Плотность распределения вероятностей обладает двумя основными свойствами:

1. *Плотность распределения неотрицательна для всех x, т. е.* $f(x) \geqslant 0$.

2. *Интеграл от плотности распределения* $f(x)$, *взятый по всему интервалу возможных значений случайной величины, равен единице, т. е.*

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Геометрическая интерпретация свойств функций $f(x)$ и $F(x)$ сводится к следующему:

а) график плотности вероятности является непрерывной кривой, и в силу неотрицательности плотности эта кривая лежит выше оси абсцисс;

б) полная площадь, ограниченная кривой, равна единице;

в) сама кривая асимптотически приближается к оси абсцисс при условии, что случайная величина может принимать все значения числовой оси или все значения полупрямой.

В тех случаях, когда расстояние между значениями дискретной случайной величины (интервал, единица измерения) достаточно мало по сравнению с самой наблюдаемой величиной, бывает удобно рассматривать данную величину в качестве непрерывной случайной величины. Такой подход может заметно упростить решение некоторых лингвистических задач.

Примером может служить использование непрерывной *функции Вейбулла* для вычисления дискретных накопленных вероятностей в частотном списке слов, словоформ и словосочетаний [6, с. 136, 137].

Исходное выражение функции Вейбулла имеет вид

$$p_i^* = 1 - e^{-ci^k}, \qquad (6.22)$$

где p_i^* — накопленная вероятность i-й словоформы, c и k — коэффициенты распределения. Значения этих коэффициентов для различных частотных списков показаны в табл. 6.5.

Таблица 6.5

Значения коэффициентов c и k относительно некоторых языков и стилей

№	Язык или подъязык	Единицы подсчета	c	k
1	Русские деловые тексты	Словоформы	0,05357	0,44640
2	Русские деловые тексты	Основы	0,07057	0,48440
3	Английские корабельные тексты	Словоформы	0,15532	0,33040
4	Английские тексты по сельхозтехнике	Именные трехсловные сочетания	0,01815	0,41556
5	Английские тексты по электронике	Трехсловные сочетания	0,01244	0,41062
6	Английские тексты по электронике	Именные трехсловные сочетания	0,01076	0,47471
7	Немецкие публицистические тексты	Словоформы	0,15534	0,28850
8	Немецкие публицистические тексты	Трехсловные сочетания	0,00585	0,45486
9	Немецкие публицистические тексты	Именные трехсловные сочетания	0,02813	0,39088
10	Немецкие тексты по электронике	Трехсловные сочетания	0,01355	0,38860
11	Румынский и молдавский	Слова	0,17331	0,29540
12	Румынские публицистические тексты	Трехсловные сочетания	0,00594	0,55741

Используя выражение (6.22), можно получить также значения вероятностей лингвистических единиц, находящихся на i-м месте в частотном списке. К этим значениям можно прийти либо через плотность распределения

$$p_i^{*\prime} = (1 - e^{-ci^k})' = cki^{k-1}e^{-ci^k} = p_i, \qquad (6.23)$$

либо с помощью равенства

$$p_i = p_{i-1}^* - p_i^* = e^{-c(i-1)^k} - e^{-ci^k}. \qquad (6.24)$$

Так, например, применяя соотношение (6.23) и опираясь на данные табл. 6.5, нетрудно показать, что вероятность словоформ с $i = 10$ в частотном списке словоформ из русских деловых текстов составляет

$$p_{10} = 0{,}05357 \cdot 0{,}4464 \cdot 10^{(0{,}4464-1)} \cdot 2{,}7183^{-0{,}05357 \cdot 10^{0{,}4464}} \approx 0{,}006.$$

К аналогичному результату можно прийти, используя соотношение (6.24):

$$p_{10} = 2{,}7183^{-0{,}05357 \cdot 9^{0{,}4464}} - 2{,}7183^{-0{,}05357 \cdot 10^{0{,}4464}} =$$
$$= 0{,}8670 - 0{,}8602 = 0{,}0068 \approx 0{,}006.$$

8. Характеристики распределения случайной величины. Функция распределения, ряд распределения и плотность распределения наиболее полно и исчерпывающе характеризует дискретную или непрерывную случайную величину. Однако расчеты, связанные с определением этих характеристик, весьма сложны и громоздки.

Обычно масса вероятности случайной величины X сосредоточена в большей своей части внутри относительно узкого интервала значений случайной величины (см. рис. 39 и 40). Определив положение и характер этого интервала, мы получаем достаточно точное представление о распределении в целом. Эту задачу можно решить, используя усредненные числовые характеристики распределения — моменты, которые в сжатой, компактной форме указывают на наиболее существенные свойства распределения. Начнем рассмотрение этих характеристик с математического ожидания.

Математическое ожидание указывает на центр группировки значений случайной величины. Для дискретной случайной величины *математическим ожиданием называется сумма произведений всех возможных значений случайной величины на их вероятности*:

$$M(X) = x_0 p_0 + x_1 p_1 + x_2 p_2 + \ldots + x_N p_N = \sum_{i=0}^{N} x_i p_i. \quad (6.25)$$

Понятие математического ожидания распространяется и на непрерывные величины: *математическим ожиданием непрерывной случайной величины называется интеграл от произведения ее значений x на плотность распределения вероятностей $f(x)$*, т. е.

$$M(X) = \int_{-\infty}^{\infty} x f(x)\, dx. \quad (6.26)$$

Математическое ожидание обладает следующими свойствами, имеющими принципиальное значение при решении лингвистических и инженерно-лингвистических задач:

1. *Математическое ожидание постоянной (неслучайной) величины C равно ей самой*, т. е.

$$M(C) = C. \quad (6.27)$$

2. *Математическое ожидание произведения постоянной величины C на случайную величину X равно произведению постоянной на математическое ожидание этой случайной величины*, т. е.

$$M(CX) = CM(X). \quad (6.28)$$

Из свойств 1 и 2 вытекают два следствия:

а) *математическое ожидание суммы постоянной величины и случайной величины равно сумме постоянной C и математического ожидания случайной величины $M(X)$*, т. е.

$$M(C + X) = C + M(X); \quad (6.29)$$

б) *математическое ожидание линейной функции $Y = b + aX$ равно сумме постоянной b и произведения постоянной a на математическое ожидание случайной величины X*, т. е.

$$M(Y) = M(b + aX) = b + aM(X). \quad (6.30)$$

3. *Математическое ожидание суммы случайных величин X_1, X_2, X_3, ..., X_n равно сумме их математических ожиданий*, т. е.

$$M(X_1 + X_2 + \ldots + X_n) = M(X_1) + M(X_2) + \ldots + M(X_n). \quad (6.31)$$

Из этого свойства вытекает такое следствие: *математическое ожидание разности случайных величин равно разности их математических ожиданий*, т. е.

$$M(X - Y) = M(X) - M(Y). \quad (6.32)$$

4. *Математическое ожидание попарно независимых случайных величин X_1, X_2, ..., X_n равно произведению их математических ожиданий*, т. е.

$$M(X_1 X_2 \ldots X_n) = M(X_1)\, M(X_2) \ldots M(X_n). \quad (6.33)$$

Разброс возможных значений случайной величины вокруг ее центра — математического ожидания — характеризуется *теоретической дисперсией* (или просто *дисперсией*), которую можно определить как *математическое ожидание квадрата отклонения случайной величины от ее математического ожидания*:

$$D(X) = M[X - M(X)]^2. \quad (6.34)$$

Теоретическая дисперсия для дискретной случайной величины вычисляется по формуле

$$D(x) = \sigma^2 = \sum_i p_i [x_i - M(X)]^2. \quad (6.35)$$

Для непрерывной случайной величины дисперсия равна

$$D(X) = \sigma^2 = \int_{-\infty}^{+\infty} [x - M(X)]^2 f(x)\, dx. \quad (6.36)$$

Теоретическая дисперсия имеет размерность квадрата случайной величины. Между тем из соображений наглядности в оценке рассеивания удобнее пользоваться величиной, размерность которой совпадает с размерностью случайной величины X. Это достигается

путем извлечения квадратного корня из дисперсии, в результате чего получается *среднее квадратическое отклонение*:

$$\sigma = \sqrt{D(X)}. \qquad (6.37)$$

Величина σ имеет ту же размерность, что и математическое ожидание случайной величины. Это дает возможность оценивать степень вариации в данном распределении с помощью *коэффициента вариации*:

$$V(\sigma) = \frac{\sigma}{M(X)} 100\,\%. \qquad (6.38)$$

Дисперсия характеризуется следующими свойствами:

1. *Дисперсия постоянной равна нулю*, т. е.

$$D(C) = 0. \qquad (6.39)$$

2. *Постоянную можно вынести за знак дисперсии, возведя ее в квадрат*, т. е.

$$D(CX) = C^2 D(X). \qquad (6.40)$$

3. *Увеличение (уменьшение) случайной величины на одну и ту же постоянную величину C не изменяет дисперсии*, т. е.

$$D(X + C) = D(X). \qquad (6.41)$$

4. *Дисперсия случайной величины равна математическому ожиданию квадрата случайной величины без квадрата ее математического ожидания*, т. е.

$$D(X) = M(X^2) - [M(X)]^2. \qquad (6.42)$$

5. *Дисперсия суммы конечного числа попарно независимых случайных величин равна сумме их дисперсий*, т. е.

$$D(X_1 + X_2 + \ldots + X_n) = D(X_1) + D(X_2) + \ldots + D(X_n). \qquad (6.43)$$

Дисперсия и математическое ожидание, с которыми мы только что познакомились, являются частными случаями понятия м о м е н т а, т. е. характеристики вида распределения.

Существует несколько видов моментов.

Моментом h-го порядка случайной величины X называется математическое ожидание h-й степени отклонений случайной величины от постоянной a, выступающей в качестве произвольно фиксированного начала отсчета, так называемого условного нуля.

При решении вероятностных лингвистических задач обычно используются моменты первых пяти порядков ($h = 0, 1, 2, 3, 4$):

$\nu'_0(a) = M(X-a)^0 = 1$ (*момент нулевого порядка*);
$\nu'_1(a) = M(X-a)$ (» *первого* »);
$\nu'_2(a) = M(X-a)^2$ (» *второго* »);
$\nu'_3(a) = M(X-a)^3$ (» *третьего* »);
$\nu'_4(a) = M(X-a)^4$ (» *четвертого* »).

Если $a = 0$, то момент называется *начальным* и записывается в общем виде так:

$$\nu_h = M(X^h).$$

Запишем выражения для начальных моментов первых пяти порядков:

$\nu_0 = M(X^0) = 1$ (*начальный момент нулевого порядка*);
$\nu_1 = M(X)$ (» » *первого* »);
$\nu_2 = M(X^2)$ (» » *второго* »);
$\nu_3 = M(X^3)$ (» » *третьего* »);
$\nu_4 = M(X^4)$ (» » *четвертого* »).

Нетрудно заметить, что начальный момент первого порядка есть не что иное, как математическое ожидание случайной величины X.

Если в качестве начала отсчета взято математическое ожидание случайной величины, т. е. если $a = M(X)$, то момент называется *центральным*. В общем виде центральный момент записывается так:

$$\mu_h = M[X - M(X)]^h. \qquad (6.44)$$

Между центральными и начальными моментами существует прямая связь, передаваемая следующей зависимостью:

$$\mu_h = \sum_{d=2}^{h} (-1)^{h-d} C_n^d \nu_1^{h-d} \nu_d + (-1)^{h-1} (h-1) \nu_1^h. \qquad (6.45)$$

Из соотношения (6.45) получаем выражения для первых пяти центральных моментов:

$\mu_0 = 1$ (*центральный момент нулевого порядка*);
$\mu_1 = 0$ (« « *первого* «);
$\mu_2 = \nu_2 - \nu_1^2$ (« « *второго* «);
$\mu_3 = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3$ (« « *третьего* «);
$\mu_4 = \nu_4 - 4\nu_3\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4$ (« « *четвертого* «).

Центральный момент первого порядка, равный нулю, представляет собой математическое ожидание случайной величины X при условии, что $a = M(X)$.

Центральный момент второго порядка является дисперсией случайной величины:

$$\mu_2 = \nu_2 - \nu_1^2 = M(X^2) - [M(X)]^2 = D(X) = \sigma^2. \qquad (6.46)$$

Отношение центрального момента h-го порядка к h-й степени среднего квадратического отклонения σ называется *нормированным моментом*. В общем виде нормированный момент записывается следующим образом:

$$\rho_h = \mu_h / (\sqrt{\mu_2})^h = \mu_h / \sigma^h. \qquad 6.47)$$

Для нормированных моментов первых четырех порядков имеем:

$\rho_1 = 0$ (*нормированный момент первого порядка*);
$\rho_2 = 1$ (« « *второго* «);
$\rho_3 = \mu_3/\sigma^3$ (« « *третьего* «);
$\rho_4 = \mu_4/\sigma^4$ (« « *четвертого* «).

Нормированный момент третьего порядка, называющийся иногда *коэффициентом асимметрии*, характеризует «скошенность» распределения.

Если распределение симметрично относительно $a = M(X)$, то центральный момент третьего порядка μ_3 (как и вообще все центральные моменты нечетных порядков) равен нулю, в связи с чем коэффициент асимметрии $\rho_3 = \mu_3/\sigma^3 = 0$.

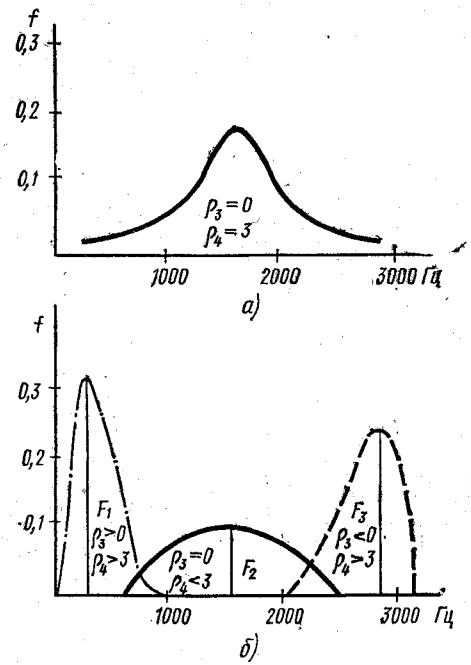


Рис. 42

Если же $\rho_3 > 0$, то график распределения характеризуется правосторонней (положительной) скошенностью. При $\rho_3 < 0$ имеем левостороннюю (отрицательную) скошенность.

Четвертый нормированный момент используется в качестве характеристики «крутизны» (островершинности или плосковершинности). Для наиболее важного с точки зрения лингвистики распределения — нормального распределения (см. ниже, § 3, п. 4) — выполняется равенство

$$\rho_4 = \mu_4/\sigma^4 = 3. \quad (6.48)$$

Для распределений, отличающихся от нормального более острой вершиной, $\rho_4 > 3$; напротив, распределения с более плоскими вершинами (или имеющие «выемку» в центре) дают $\rho_4 < 3$.

Понятия «скошенности» и «крутизны» иллюстрирует рис. 42. На рис. 42, *а* изображена эталонная кривая нормального распределения, а на рис. 42, *б* — кривые распределения относительных частот значений (в Гц) первых трех формант (F_1, F_2, F_3) русских гласных.

Наряду с четвертым нормированным моментом для измерения крутизны распределения используется величина

$$E = \rho_4 - 3, \quad (6.49)$$

называемая *эксцессом* (или *куртозисом*).

Наиболее важными при описании распределений, использующихся в языкознании, являются следующие моменты: $\nu_1 = M(X) = a$, $\mu_2 = D(X) = \sigma^2$, ρ_3, ρ_4.

§ 3. Законы распределения, моделирующие образование языковых единиц текста

В теории вероятностей известны десятки законов распределения случайной величины. Задача квантитативной лингвистики состоит в том, чтобы найти среди них такие законы, которые могли бы выступать в качестве наиболее адекватных математических моделей порождения текста и составляющих его языковых единиц.

1. Биномиальное распределение. Исходной схемой при построении многих законов распределений является уже знакомое нам биномиальное распределение

$$P(x) = C_N^x p^x q^{N-x},$$

которое характеризуется следующими параметрами:

а) математическим ожиданием (начальным моментом первого порядка)

$$M(X) = Np; \quad (6.50)$$

б) дисперсией (центральным моментом второго порядка)

$$D(X) = Npq \quad (6.51)$$

и средним квадратическим отклонением

$$\sigma = \sqrt{Npq}; \quad (6.52)$$

в) коэффициентом асимметрии

$$\rho_3 = (q-p)/\sqrt{Npq}; \quad (6.53)$$

г) четвертым нормированным моментом

$$\rho_4 = (1 - 6pq + 3Npq)/(Npq) \quad (6.54)$$

и соответственно эксцессом

$$E = \rho_4 - 3 = (1 - 6pq)/(Npq). \quad (6.55)$$

Если частость $f = X/N$, так же как и случайная величина, распределена по биномиальному закону, то значения математического ожидания, дисперсии и среднего квадратического отклонения таковы:

$$M\left(\frac{X}{N}\right) = \frac{1}{N} M(X) = \frac{1}{N} \cdot Np = p, \quad (6.56)$$

$$D\left(\frac{X}{N}\right) = \frac{1}{N^2} D(X) = \frac{1}{N^2} \cdot Npq = \frac{pq}{N}, \quad (6.57)$$

$$\sigma\left(\frac{X}{N}\right) = \sigma = \sqrt{\frac{pq}{N}}. \quad (6.58)$$

Как уже говорилось, это распределение может быть использовано при описании употребления фонем, графем и их классов, а также грамматических категорий при условии, что величины N и x не очень велики. Однако в конкретных лингвистических задачах эти условия обычно не соблюдаются, поэтому вместо биномиального распределения приходится использовать другие аппроксимирующие его распределения.

2. Распределение редких лингвистических единиц (распределение Пуассона). Будем рассматривать последовательные появления интересующей нас языковой единицы A в тексте T в качестве *потока лингвистических событий*. Примерами такого потока могут служить последовательные появления в русском связном тексте слова *море* в различных его формах, или словоформы *моря*, или словосочетания *у самого синего моря* и т. п.

Поток лингвистических событий называется *простейшим* в том случае, когда выполняются следующие условия:

1) если разбить текст T на N отрезков равной длины, то вероятность появления лингвистических событий в отрезке $t = T/N$ зависит только от длины этого отрезка (но не от начала отсчета). Это условие позволяет оперировать вероятностью $P(x)$ того, что в любом отрезке t лингвистическая единица A появится ровно x раз ($x = 0, 1, 2, \ldots$);

2) вероятность появления лингвистической единицы A практически не зависит от того, сколько раз употреблялась единица A до этого в тексте и употреблялась ли она вообще. Такое предположение оказывается вполне корректным для редких лингвистических единиц, поскольку валентности этих единиц распространяются не более чем на семь значащих шагов (обычно словоформ) текста [23, с.58];

3) вероятность наступления двух и более лингвистических событий в бесконечно малом отрезке текста t есть бесконечно малая величина более высокого порядка, чем величина t. Например, разбив текст на равные отрезки длиной в десять букв каждый, мы можем считать ничтожно малой вероятностью появление в одном отрезке двух словоформ *моря*;

4) при дальнейшем уменьшении отрезка t вероятность наступления одного лингвистического события убывает пропорционально длине t. Таким образом, чем меньше интервал, тем меньше, вследствие малости интервала, вероятность появления лингвистической единицы A.

Из всего сказанного следует, что неограниченное уменьшение p и t пропорционально неограниченному увеличению N. В связи с этим произведение вероятности p и числа отрезков N (объема выборки), представляющее собой математическое ожидание случайной величины, является постоянным:

$$M(X) = Np = \lambda.$$

Отсюда $p = \lambda/N$. Подставляя это значение в формулу биномиальной вероятности [см. соотношение (6.1)], получим

$$P_N(x) = C_N^x \left(\frac{\lambda}{N}\right)^x \left(1-\frac{\lambda}{N}\right)^{N-x} =$$

$$= \frac{\lambda^x}{x!}\left(1-\frac{\lambda}{N}\right)^N \frac{N}{N}\frac{N-1}{N}\cdots\frac{N-x+1}{N}\frac{1}{\left(1-\frac{\lambda}{N}\right)^x},$$

где

$$\lim_{N\to\infty}\left(1-\frac{\lambda}{N}\right)^N = e^{-\lambda}, \quad \lim_{N\to\infty}\frac{1}{\left(1-\frac{\lambda}{N}\right)^x} = 1,$$

а предел каждого члена $\frac{N-m}{N}$ ($m = 0, 1, \ldots, x-1$) составляет $\lim\limits_{N\to\infty}\frac{N-m}{N} = 1$.

В итоге получаем

$$\lim_{N\to\infty} P_N(x) = \frac{\lambda^x}{x!}e^{-\lambda},$$

или

$$P(x, \lambda) = \frac{\lambda^x}{x!}e^{-\lambda}. \qquad (6.59)$$

Это и есть **формула распределения Пуассона**, использующаяся для описания употребления редких лингвистических единиц. Единственным параметром этого распределения является величина λ. В лингвистических приложениях λ есть среднее число употреблений интересующего нас языкового элемента в тексте. Аргумент x — обычно это число употреблений лингвистической единицы — принимает значения 0, 1, 2, ...

Аналитическое выражение распределения вероятностей $P(x, \lambda)$ показано в табл. 6.6.

Таблица 6.6

x	0	1	2	...	x	...	N	...
$P(x, \lambda)$	$e^{-\lambda}$	$\lambda e^{-\lambda}$	$\frac{\lambda^2}{2}e^{-\lambda}$	...	$\frac{\lambda^x}{x!}e^{-\lambda}$	...	$\frac{\lambda^N}{N!}e^{-\lambda}$	...

Распределение Пуассона характеризуется следующими параметрами:

а) математическим ожиданием (начальным моментом первого порядка), дисперсией (центральным моментом второго порядка), а также центральным моментом третьего порядка, равными λ:

$$M(X) = D(X) = \mu_3 = \lambda, \qquad (6.60)$$

и среднеквадратическим отклонением

$$\sigma = \sqrt{M(X)} = \sqrt{\lambda}; \qquad (6.61)$$

б) коэффициентом асимметрии

$$\rho_3 = \lambda/(\sqrt{\lambda})^3 = 1/\sqrt{\lambda}; \tag{6.62}$$

в) четвертым нормированным моментом

$$\rho_4 = (3\lambda + 1)/\lambda \tag{6.63}$$

и эксцессом, равным

$$E = \frac{\lambda(3\lambda+1)}{\lambda^2} - 3 = \frac{1}{\lambda}. \tag{6.64}$$

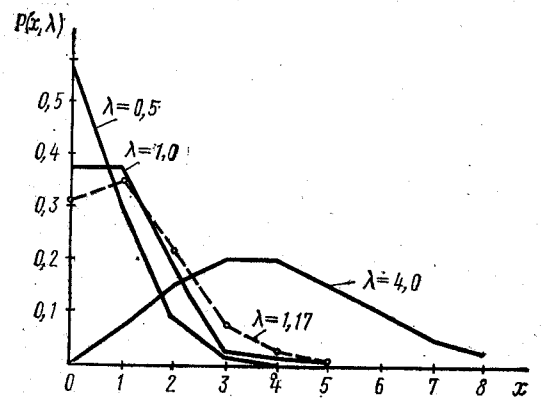


Рис. 43. Распределение Пуассона при различных значениях параметра λ:
——— кривые распределения Пуассона при λ=0,5; 1,0; 4,0; ○——○——○ кривая распределения вероятностей появления в немецком тексте сегмента Δdaβ die (λ=1,17)

При малых значениях λ распределение сосредоточено вблизи начала координат. С ростом λ оно приобретает колоколообразную форму с правосторонней скошенностью (рис. 43). При больших значениях параметра λ ($\lambda \geqslant 10$) распределение Пуассона приобретает форму нормального распределения при $M(X) = \lambda$ и $D(X) = \sigma^2 = \lambda$.

Как и при биномиальном распределении, вычисление вероятностей пуассоновского распределения удобно начинать с определения моды x_0 и соответствующей ему модальной вероятности $P(x_0, \lambda)$. Рассмотрим в этой связи поведение аргумента x в нашем распределении. Из равенства

$$\frac{P(x,\ \lambda)}{P(x-1,\ \lambda)} = \frac{\lambda^x e^{-\lambda}/x!}{\lambda^{x-1} e^{-\lambda}/(x-1)!} = \frac{\lambda}{x} \tag{6.65}$$

видно, что если $x > \lambda$, то $P(x, \lambda) < P(x-1, \lambda)$; если же $x < \lambda$, то $P(x, \lambda) > P(x-1, \lambda)$; наконец, если $x = \lambda$, то $P(x, \lambda) = P(x-1, \lambda)$. Ясно также, что $P(x, \lambda)$ возрастает при увеличении x от нуля до $x_0 = [\lambda]$ и* убывает при дальнейшем росте x. В том случае, когда λ является целым числом, $P(x, \lambda)$ имеет два модальных значения: при $x_0 = \lambda$ и при $x_0' = \lambda - 1$; в том же случае, когда λ — дробное число, $P(x, \lambda)$ имеет одно модальное значение при $x_0 = [\lambda]$.

* Символом [λ] обозначена целая часть значения λ.

Определив таким образом модальное значение $x_0 = [\lambda]$, найдем его вероятность по формуле

$$P(x_0,\ \lambda) = \frac{\lambda^{x_0}}{x_0!} e^{-\lambda} = \frac{\lambda^{[\lambda]}}{[\lambda]!} e^{-\lambda}. \tag{6.66}$$

Вычисление остальных вероятностей осуществляется по рекуррентным формулам, вытекающим из (6.65):

при $x < x_0$

$$\left.\begin{aligned} P(x_0-1,\ \lambda) &= \frac{x_0}{\lambda} P(x_0,\ \lambda), \\ P(x_0-2,\ \lambda) &= \frac{x_0-1}{\lambda} P(x_0-1,\ \lambda), \\ &\cdots\cdots\cdots\cdots\cdots\cdots \end{aligned}\right\} \tag{6.67а}$$

при $x > x_0$

$$\left.\begin{aligned} P(x_0+1,\ \lambda) &= \frac{\lambda}{x_0+1} P(x_0,\ \lambda), \\ P(x_0+2,\ \lambda) &= \frac{\lambda}{x_0+2} P(x_0+1,\ \lambda). \\ &\cdots\cdots\cdots\cdots\cdots\cdots \end{aligned}\right\} \tag{6.67б}$$

В лингвистике, как уже говорилось, используется бесповторное статистическое исследование, при котором текст разбивается на S выборок (серий), каждая длиной в N лингвистических элементов. При этом бывает необходимо определить теоретически ожидаемое число серий S_x^T, в которых лингвистический элемент A появится ровно x раз.

Пользуясь рассуждениями, приведенными в п. 3 § 1 относительно биномиального распределения, нетрудно показать, что

$$S_x^T = SP(x,\ \lambda) = S\,\frac{\lambda^x}{x!} e^{-\lambda}, \tag{6.68}$$

где S выступает в качестве коэффициента пропорциональности, связывающего величины S_x^T и $P(x, \lambda)$, а

$$\lambda \simeq \bar{x} = \frac{1}{S} \Sigma x S_x.$$

Рассмотрим в связи с этим следующий пример. В немецких публицистических текстах осуществлена выборка 100 серий по 1000 трехсловных сегментов. Используя данные, приведенные в табл. 6.7, определим теоретическое пуассоновское распределение вероятностей появления сегментов Δdaβ der (начало дополнительного придаточного предложения с существительным в начальной позиции) в одной серии и теоретическое количество серий S_x^T, содержащих по 0, 1, 2,... таких сегментов.

Таблица 6.7

Пуассоновское распределение вероятностей появления сегмента Δ daβ der в немецких публицистических текстах

Число появлений события x	Эмпирические частоты появления выборок S_x	Теоретически ожидаемое число выборок S_x^T	S_x^T округленное до целых чисел	Частость $\frac{S_x}{S}$	Вероятность $P(x, \lambda)$
0	32	31,04	31	0,32	0,3103
1	38	36,28	36	0,38	0,3628
2	19	21,24	21	0,19	0,2123
3	5	8,28	8	0,05	0,0828
4	4	2,42	3	0,04	0,0242
5	2	0,57	1	0,02	0,0057
Суммы	100		100	1,00	0,9981

Сначала вычислим значение λ:

$$\lambda \approx \bar{x} = \frac{1}{100}(0\cdot 32 + 1\cdot 38 + 2\cdot 19 + 3\cdot 5 + 4\cdot 4 + 5\cdot 2) =$$

$$= \frac{1}{100}(38 + 38 + 15 + 16 + 10) = \frac{117}{100} = 1{,}17.$$

Модальное значение x_0 равно целой части значения λ, т. е. единице. Значение $P(x_0, \lambda)$ будем искать с помощью табл. I, помещенной в Приложении (см. стр. 362, 363). В этой таблице значения вероятности, соответствующей $\lambda = 1{,}17$ и $x = 1$, нет. Поэтому обращаемся к интерполяции. Для этого берем значения $P(1;\ 1{,}1) = 0{,}36616$ и $P(1;\ 1{,}2) = 0{,}36143$; тогда искомая вероятность

$$P(1;\ 1{,}17) = 0{,}36143 + \frac{0{,}36616 - 0{,}36143}{10}\cdot 3 =$$

$$= 0{,}36143 + \frac{0{,}00473}{10}\cdot 3 = 0{,}3628.$$

После этого, используя рекуррентные формулы (6.67), так, как это показано в п. 3 § 1, находим по табл. I остальные значения $P(x, \lambda)$. Затем, с помощью выражения (6.68) получаем значения S_x^T, которые также приведены в нашей таблице.

Степень близости теоретического и эмпирического распределений, приведенных в табл. 6.7, мы рассмотрим ниже.

В разделе, посвященном биномиальному распределению, уже говорилось о том, что как с прикладной, так и с теоретико-языковедческой точки зрения важно уметь определять вероятность появления лингвистического элемента от a до b раз. В этом случае имеем

$$P(a \leqslant x \leqslant b,\ \lambda) = \sum_{x=a}^{b} P(x,\ \lambda) = \sum_{x=a}^{b} \frac{\lambda^x}{x!} e^{-\lambda}.$$

Весь ход решения этой задачи и ее частные случаи аналогичны процессу решения, описанному в п. 7 § 1 с той лишь разницей, что биномиальные вероятности заменяются пуассоновскими.

Таким же способом определяется вероятность появления редкого лингвистического события хотя бы один раз. Она равна

$$P(x \geqslant 1,\ \lambda) = 1 - P(x = 0,\ \lambda) = 1 - e^{\lambda}. \qquad (6.69)$$

Отсюда по схеме, описанной в п. 8 § 1, рассчитывается объем выборки N, необходимый для того, чтобы обеспечить с заданной вероятностью появление хотя бы один раз определенного лингвистического элемента.

С этой целью приведем выражение (6.69) к виду

$$e^{Np} = 1 - P(x \geqslant 1,\ \lambda),$$

а затем, прологарифмировав обе части и произведя необходимые преобразования, получим

$$N = \frac{\lg[1 - P(x > 1,\ \lambda)]}{p \lg e}. \qquad (6.70)$$

Используя приведенные выше числовые характеристики словоформы *напряжение*, определим объем текста, необходимый для того, чтобы указанная словоформа с вероятностью в 90% появилась в нем хотя бы один раз (предполагается, что в данном случае имеет место распределение Пуассона).

Применяя соотношение (6.70), находим

$$N = \frac{\lg(1 - 0{,}90)}{0{,}0023 \lg e} = \frac{\lg 0{,}10}{-0{,}0023\cdot 0{,}44} \approx \frac{-1}{-0{,}001} = 1000.$$

Заметим, что объем необходимой выборки здесь тот же, что и в том случае, когда мы предполагали, что словоформа *напряжение* имеет биномиальное распределение (ср. § 1, п. 8).

3. **Распределения, описывающие взаимодействие случайных и детерминированных процессов в речи (распределения Чебанова — Фукса и Фукса — Гачечиладзе).** Применяя биномиальное и пуассоновское распределение для исследования поведения дискретных лингвистических единиц в речи, мы исходили из предположения, что речь представляет собой простейший поток лингвистических событий. Однако этот подход, представляющий собой очень грубую и упрощенную аппроксимацию лингвистических явлений, имеет ограниченное применение в лингвистике. Выше уже говорилось (см. гл. 5), что формирование слов, словосочетаний, предложений, высказываний представляет собой взаимодействие как случайных, так и детерминированных процессов. Случайными с лингвистической точки зрения являются описываемые текстом ситуации объективной действительности, детерминированными же представляются некоторые правила системы и нормы языка. Поэтому для описания образования лингвистических единиц и их распределения в тексте следует

применять такие распределения, которые учитывали бы взаимодействие этих случайных процессов и детерминированных правил.

Одним из простейших распределений, описывающих лингвостатистическую структуру, возникающую из комбинации абсолютно случайного и абсолютно детерминированного процесса, является распределение Чебанова — Фукса.

Это распределение строится исходя из следующих рассуждений. Наряду с *бескомпонентными* лингвистическими величинами, т. е. с такими дискретными величинами, которые могут принимать любое положительное целочисленное значение, а также значение нуль [т. е. $P(0) \geqslant 0$], в языке в основном функционируют *небескомпонентные* величины, которые согласно законам системы языка никогда не могут принять значение нуль. К таким величинам относятся, например, длины слогов (силлабографов), морфем, слов, предложений и т. п. Действительно, каждый слог состоит, по крайней мере, из одной фонемы, а каждое слово состоит, по крайней мере, из одного слога или морфемы [здесь $P(0) \neq 0$]. Таким образом, небескомпонентная случайная лингвистическая величина может принимать значения 1, 2, 3, ...

Если небескомпонентная случайная величина встречается достаточно редко в тексте, то распределение вероятностей ее значений в тексте может быть описано с помощью распределения Пуассона при условии, что заданный системой языка обязательный элемент будет исключен из рассмотрения. Иными словами, вместо аргумента x следует взять разность $x - 1$, а вместо средней λ — разность $\lambda - 1$. Учитывая эти условия, получаем равенство

$$P(x,\ \lambda-1) = \frac{(\lambda-1)^{x-1}}{(x-1)!}\, e^{-(\lambda-1)}, \qquad (6.71)$$

представляющее собой **формулу распределения Чебанова — Фукса.**

Аналитическое выражение вероятностей $P(x, \lambda - 1)$ показано в табл. 6.8.

Таблица 6.8

x	1	2	3	...	x	...	N	...
$P(x,\lambda-1)$	$e^{-(\lambda-1)}$	$(\lambda-1)\times e^{-(\lambda-1)}$	$\frac{(\lambda-1)^2}{2}\times e^{-(\lambda-1)}$	...	$\frac{(\lambda-1)^{x-1}}{(x-1)!}\times e^{-(\lambda-1)}$	...	$\frac{(\lambda-1)^{N-1}}{(N-1)!}\times e^{-(\lambda-1)}$	...

Распределение Чебанова—Фукса характеризуется следующими параметрами:

а) первым начальным моментом (математическим ожиданием)

$$\nu_1 = M(X) = \lambda; \qquad (6.72)$$

б) вторым центральным моментом (дисперсией)

$$\mu_2 = D(X) = \lambda - 1 \qquad (6.73)$$

и средним квадратическим отклонением

$$\sigma = \sqrt{\lambda - 1}; \qquad (6.74)$$

в) коэффициентом асимметрии

$$\rho_3 = 1/\sqrt{\lambda - 1}; \qquad (6.75)$$

г) четвертым нормированным моментом

$$\rho_4 = 1/(\lambda - 1) + 3, \qquad (6.76)$$

и эксцессом

$$E = 1/(\lambda - 1). \qquad (6.77)$$

Так же как и пуассоновское распределение, **распределение Чебанова—Фукса** определяется параметром λ.

Поведение этого распределения точно соответствует поведению распределения Пуассона с учетом сдвига на одну единицу: при малых значениях λ распределение имеет острую вершину и сосредоточено около прямой $x = 1$. С ростом λ оно постепенно приобретает более пологую колоколообразную форму с правосторонней скошенностью, которая уменьшается по мере роста λ (рис. 44).

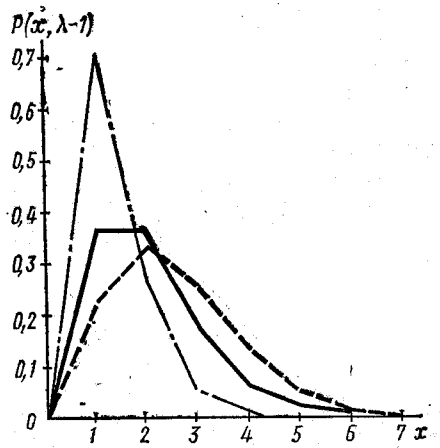


Рис. 44. Распределение Чебанова — Фукса при разных значениях параметра λ:
——— $\lambda = 2{,}00$;
——— $\lambda = 1{,}35$ (распределение вероятностей слоговых структур английского слова).
– – – – – $\lambda = 2{,}54$ (распределение вероятностей слоговых структур грузинского слова).

С помощью распределения Чебанова — Фукса делались попытки описать вероятностный процесс образования слов из слогов и морфем по различным языкам. Теоретические и эмпирические данные по распределению вероятностей слоговых длин слов показаны в табл. 6.9, а распределение вероятностей морфемных длин — в табл. 6.10.

Легко заметить, что экспериментальные результаты дают иногда отклонения от теоретических оценок, на это указывают, в частности, высокие значения сумм абсолютных величин линейных отклонений для слоговой структуры румынского и арабского слова, а также для морфемной структуры английского слова.

Эти расхождения опытных данных и теоретической модели говорят о том, что распределение Чебанова—Фукса нельзя рассматривать в качестве универсального закона, описывающего основные свойства процесса образования лингвистических единиц.

Таблица 6.9

Распределение вероятностей длин слогов в некоторых языках мира*

$P(x)$ \ Языки	Английский		Немецкий		Румынский		Арабский		Древнегреческий		Русский		Грузинский	
	Теория	Эксперим.	Теория	Эксперим.	Теория	Эксперим.	Теория	Эксперим.	Теория	Эксперим.	Теория	Эксперим.	Теория	Эксперим.
P (1)	0,705	0,715	0,532	0,556	0,387	0,468	0,332	0,227	0,332	0,376	0,292	0,339	0,214	0,238
P (2)	0,246	0,194	0,336	0,308	0,368	0,272	0,365	0,497	0,365	0,321	0,359	0,303	0,331	0,312
P (3)	0,043	0,068	0,106	0,094	0,175	0,142	0,201	0,224	0,201	0,168	0,221	0,214	0,255	0,224
P (4)	0,005	0,016	0,022	0,033	0,055	0,077	0,074	0,050	0,074	0,089	0,090	0,097	0,131	0,146
P (5)	0,001	0,006	0,004	0,007	0,013	0,039	0,021	0,002	0,021	0,034	0,027	0,036	0,051	0,061
P (6)	0,000	0,001	0,001	0,002	0,002	0,002	0,005	0,000	0,005	0,008	0,007	0,010	0,016	0,015
P (7)							0,002	0,000	0,002	0,001	0,003	0,001	0,004	0,003
P (8)											0,001	0,000	0,001	0,001
λ (в слогах)	1,35		1,63		1,95		2,104		2,105		2,23		2,54	
Сумма абсолютных величин линейных отклонений	0,095		0,079		0,258		0,310		0,153		0,132		0,101	
H (дв. ед.)	1,22		1,51		1,83		1,70		2,02		2,02		2,28	

* Таблица составлена по данным, взятым из работы [23] и из следующих работ: [320, с. 113—143]; [48, с. 84].

Таблица 6.10

Распределение вероятностей морфемных структур слов в английском и русском языках*

$P(x)$ \ Языки	Английский			Русский	
	Теория		Эксперимент	Теория	Эксперимент
	Распределение Чебанова—Фукса	Распределение Фукса—Гачечиладзе			
(1)	(2)	(3)	(4)	(5)	(6)
P (1)	0,246	0,109	0,113	0,272	0,260
P (2)	0,346	0,500	0,549	0,357	0,417
P (3)	0,242	0,285	0,238	0,230	0,165
P (4)	0,112	0,084	0,090	0,100	0,118
P (5)	0,039	0,022	0,007	0,032	0,032
P (6)	0,011	0,000	0,002	0,009	0,006
P (7)	0,003	0,000	0,001	0,003	0,002
P (8)	0,001	0,000	0,000	0,001	0,000
λ (в морфемах)	2,41			2,30	
Сумма абсолютных величин линейных отклонений	0,396		0,124	0,160	
H (дв. ед.)			1,71	2,04	

* Таблица составлена по данным, взятым из работы [23].

Очевидно, при описании процессов образования лингвистических единиц нужно учитывать несколько задаваемых системой языка детерминированных правил. Об этом можно, в частности, судить по данным румынского языка, в котором поэтическая речь показывает иной спектр вероятностей для слоговых длин слова, чем это имеет место в деловом стиле (см. табл. 6.11). Этим же, вероятно, объясняется и большое расхождение экспериментальных и теоретических данных в арабском тексте. Иными словами, процесс образования лингвистических единиц характеризуется более тонкой статистической структурой, чем та, которая представлена распределением Чебанова—Фукса. Эту структуру более точно описывает о б о б щ е н н о е р а с п р е д е л е н и е Ф у к с а — Г а ч е ч и л а д з е.

Основная идея распределения Фукса—Гачечиладзе состоит в следующем. Лингвистические единицы определенного типа могут включать не только один, но и 2, 3, ..., ν обязательных элементов. Тогда, исключая из распределения ν обязательно присутствующих элементов, мы приходим к распределению вероятностей появления лингвистических единиц различной длины, которое имеет следующий вид:

$$P(x, \lambda - \nu) = \frac{(\lambda - \nu)^{x-\nu}}{(x-\nu)!} e^{-(\lambda-\nu)}. \qquad (6.78)$$

Таблица 6.11

Распределение вероятностей длин слогов в разных стилях румынского языка*

Подъязыки	Румынская поэзия (Eminescu, Luceafărul)		Румынский деловой стиль («Гражданский кодекс»)	
	Теория	Эксперимент	Теория	Эксперимент
$P(1)$	0,549	0,580	0,301	0,400
$P(2)$	0,329	0,263	0,360	0,222
$P(3)$	0,099	0,119	0,217	0,160
$P(4)$	0,019	0,034	0,086	0,158
$P(5)$	0,004	0,003	0,026	0,049
$P(6)$	0,000	0,001	0,007	0,001
$P(7)$			0,003	0,000
λ (в слогах)	1,60		2,20	
Сумма абсолютных величин линейных отклонений	0,134		0,398	
H (дв. ед.)		1,56		2,07

* Таблица составлена по данным, взятым из работы [56].

Дальнейшее рассуждение приводит к распределению еще более общего типа. Предположим, что в некотором лингвистическом классе имеются лингвистические единицы, которые включают 0, 1, 2, ..., ..., ν языковых элементов. Среди этих единиц можно выделить единицы, образованные хотя бы из одного элемента, из двух, из трех, ..., из ν и т. д. элементов.

Обозначим вероятность появления лингвистических единиц, состоящих из 0, 1, 2, 3, ... элементов, через ε_0, статистический вес единиц, образованных хотя бы из одного элемента, через ε_1, ..., вес единиц, состоящих хотя бы из ν элементов, через ε_ν и т. д. Совокупность параметров ε_ν называется *лингвистическим спектром* образования языковых единиц с суммой этих параметров

$$A = \sum_{\nu=1}^{\infty} \varepsilon_\nu.$$

Далее все множество единиц данного типа (предложений, словосочетаний, слов, морфем, слогов) разбивается соответственно особенностям их образования на классы. Статистический вес класса ν характеризуется разностью $\varepsilon_\nu - \varepsilon_{\nu+1}$, указывающей вклад каждого детерминированного правила в процесс образования данной лингвистической единицы.

Например, русское предложение может не иметь ни одного существительного (*Пошли!*), может включать одно (*Пошли в кино!*), два (*Ребята, пошли в кино!*), ..., ν существительных. В этом случае

вес предложений, не имеющих существительных, характеризуется разностью $\varepsilon_0 - \varepsilon_1$, вес предложений с одним существительным равен $\varepsilon_1 - \varepsilon_2$ и т. д. ($\varepsilon_0 > \varepsilon_1 > \varepsilon_2 > \ldots > \varepsilon_\nu > \varepsilon_{\nu+1} \ldots$).

Все эти соображения, а также математические рассуждения, изложенные в работе [32а, с. 114—118], и приводят к формуле **обобщенного распределения Фукса—Гачечиладзе**, имеющей вид:

$$P(x, \lambda - A) = e^{-(\lambda - A)} \sum_{\nu=0}^{\infty} (\varepsilon_\nu - \varepsilon_{\nu+1}) \frac{(\lambda - A)^{x-\lambda}}{(x-\nu)!} \varphi_\nu(A, \lambda, x). \tag{6.79}$$

Величина λ определяется опытным путем, методика вычисления весовых коэффициентов ε_ν через первые три момента распределения (6.79) — см. ниже — описана в [32а]. Расчет коэффициента $\varphi_\nu(A, \lambda, x)$, учитывающего влияние контекстного окружения, дан в работе [48]. Конкретные лингвистические задачи решаются при допущении, что

$$\varphi_\nu(A, \lambda, x) \approx 1.$$

Распределение Фукса—Гачечиладзе характеризуется следующими параметрами:

а) первым начальным моментом (математическим ожиданием)

$$\nu_1 = M(X) = \lambda, \tag{6.80}$$

б) вторым центральным моментом (дисперсией)

$$\mu_2 = D(X) = \lambda^2 + \lambda - \left(\sum_{k=1}^{\infty} \varepsilon_k\right)^2 + 2 \sum_{k=1}^{\infty} \varepsilon_k (k-1) \tag{6.81}$$

и средним квадратическим отклонением

$$\sigma = \sqrt{\lambda^2 + \lambda - \left(\sum_{k=1}^{\infty} \varepsilon_k\right)^2 + 2 \sum_{k=1}^{\infty} \varepsilon_k (k-1)} \tag{6.82}$$

в) третьим центральным моментом

$$\mu_3 = \lambda^3 + 3\lambda^2 + \lambda + 2\left(\sum_{k=1}^{\infty} \varepsilon_k\right)^3 + 3\left(\sum_{k=1}^{\infty} \varepsilon_k\right)^2 - 2 \sum_{k=1}^{\infty} \varepsilon_k - 3\lambda\left(\sum_{k=1}^{\infty} \varepsilon_k\right)^2 - 6\lambda \sum_{k=1}^{\infty} \varepsilon_k + 3 \sum_{k=1}^{\infty} k^2 \varepsilon_k + \left[6\left(\lambda - \sum_{k=1}^{\infty} \varepsilon_k\right) + 1\right] \sum_{k=1}^{\infty} k\varepsilon_k. \tag{6.83}$$

Выше уже говорилось о том, что распределение Чебанова—Фукса плохо моделирует процесс образования английских слов из мор-

фем. Попробуем описать этот процесс с помощью распределения Фукса—Гачечиладзе, используя при этом следующий ε-спектр:

$$\varepsilon_0 = 0,\ \varepsilon_1 = 1,\ \varepsilon_2 = 0{,}8,\ \varepsilon_3 = \varepsilon_4 = \ldots = 0.$$

Разность $\varepsilon_1 - \varepsilon_2 = 0{,}2$ характеризует вклад в процесс одноморфемных служебных слов, а разность $\varepsilon_2 - \varepsilon_3 = 0{,}8$ указывает на вклад знаменательных слов, состоящих из двух и более морфем. Средняя $\lambda = 2{,}41$, коэффициент $\varphi_\nu(A, \lambda, 1) \approx 1$.

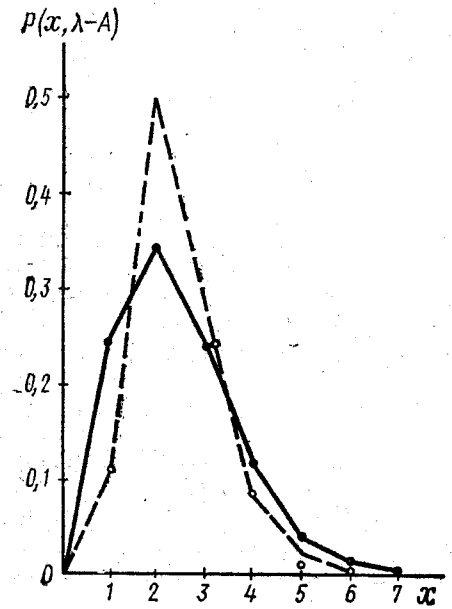


Рис. 45. Распределение вероятностей длин английского слова в морфемах:
——— теоретическая кривая распределения Чебанова — Фукса (λ=2,41);
– – – – – теоретическая кривая распределения Фукса — Гачечиладзе (λ=2,41, ε_1=1, ε_2=0,80);
○ ○ ○ ○ ○ экспериментальные точки

При этих условиях распределение вероятностей разных морфемных построений английского слова описывается следующим выражением:

$$P(x, \lambda - A) = e^{-0{,}61} \times \left[0{,}2\,\frac{0{,}61^{x-1}}{(x-1)!} + 0{,}8\,\frac{0{,}61^{x-2}}{(x-2)!}\right].$$

Данные табл. 6.10 [ср. столбцы (2), (3) и (4)], а также рис. 45 показывают, что распределение Фукса — Гачечиладзе довольно хорошо описывает распределение вероятностей различных морфемных структур английского слова. Если принять во внимание большее число членов ε-спектра, то можно, вероятно, добиться еще большей близости теоретического и эмпирического распределений.

Следует иметь в виду, что распределение Фукса — Гачечиладзе включает в себя в виде частных случаев некоторые из рассмотренных выше дискретных распределений. Действительно, полагая $\varepsilon_0 = 1$, $\varepsilon_1 = \varepsilon_2 = \ldots = \varepsilon_\nu = 0$, т. е. считая, что рассматриваемый тип лингвистических единиц однороден и не распадается на классы, а также учитывая тот факт, что образующие этот тип единицы принадлежат к бескомпонентному типу, т. е. могут включать 0, 1, 2, ... элементов, в связи с чем $A = 0$, $\nu = 0$, $\varphi_\nu(A, \lambda, x) \to 1$, мы приходим к распределению Пуассона (6.59):

$$P(x, \lambda - 0) = e^{-(\lambda - 0)} \cdot 1 \cdot \frac{(\lambda - 0)^{x-0}}{(x-0)!} \cdot 1 = \frac{\lambda^x}{x!} e^{-\lambda} = P(x, \lambda).$$

Если же речь идет о небескомпонентных единицах [при условии, что $\varepsilon_0 = 0$, $\varepsilon_1 = 1$, $\varepsilon_2 = \varepsilon_3 = \ldots = \varepsilon_\nu = 0$, откуда $A = 1$, $\nu = 1$, $\varphi_\nu(A, \lambda, x) \to 1$], то (6.79) превращается в распределение Чебанова—Фукса (6.71).

4. Распределение непрерывных случайных лингвистических величин. Нормальное распределение. До сих пор мы рассматривали распределения, описывавшие поведение в тексте дискретных случайных величин. В том случае, когда речь идет о лингвистических процессах и явлениях, осуществляющихся в субстанции выражения и субстанции содержания (см. гл. 1, § 3, п. 4), приходится иметь дело с непрерывными случайными величинами, к которым только что рассмотренные законы применены быть не могут. Распределение этих величин описывается специальными законами, среди которых наиболее важным является нормальное распределение (иначе нормальный закон, или **закон Гаусса**).

Нормальное распределение выступает в качестве предельного закона, к которому при определенных условиях приближаются другие теоретические распределения. Рассмотрим в этой связи соотношение биномиального и нормального распределений. Выше было показано (см. § 2, п. 5), что при бесконечном уменьшении интервалов в гистограмме, изображающей закон распределения случайных величин (например, высот основных формант речевого звука или чувствительности слуха относительно разных уровней интенсивности звука), ломаная линия, ограничивающая сверху площадь гистограммы, постепенно превращается в плавную кривую. Аналогичным образом полигон биномиального распределения при бесконечном увеличении объема выборки N и числа отдельных вероятностей $P_N(x)$ приближается к плавной кривой. Это геометрическое представление интерпретируется с помощью **локальной теоремы Муавра—Лапласа**: *если вероятность появления события A в каждом из независимых испытаний постоянна и равна p (где $0 < p < 1$), а $N \to \infty$, то произведение $\sqrt{Npq}$ на вероятность $P_N(x)$ появления события в этих испытаниях ровно x раз стремится к выражению $\frac{1}{\sqrt{2\pi}} e^{-z^2/2}$, где*

$$z = \frac{x - Np}{\sqrt{Npq}}. \qquad (6.84)$$

Эту замену производят в целях упрощения расчетов в прикладных задачах.

Так как в реальных лингвистических задачах число N хотя и велико, но всегда ограничено, то на основании только что сформулированной теоремы можно записать приближенное равенство

$$\sqrt{Npq}\ P_N(x) \approx \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

или

$$P_N(x) \approx \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2} = P'_N(x), \qquad (6.85)$$

где $\sigma = \sqrt{Npq}$ [14, с.74, и сл.].

Исходя из определения предела (см. гл. 2, § 1, п. 2), можно утверждать, что погрешность этого равенства есть бесконечно малая величина, стремящаяся к нулю при неограниченном возрастании N. Поэтому равенство (6.85) и называют **асимптотической формулой биномиального закона.** Использование приближенной формулы (6.85) для оценки вероятности $P_N(x)$ значительно упрощает вычисление, которое ведется здесь по следующей схеме:

1) вычисляют значения $Np = \mu$, $\sqrt{Npq} = \sigma$, $1/\sigma$;

2) по формуле (6.84) определяют значение z;

3) с помощью табл. II, помещенной в Приложении (см. стр. 364), по значению z находят

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}; \qquad (6.86)$$

4) умножив полученное значение $\varphi(z)$ на $1/\sigma$, получают значение $P'_N(x)$.

Считая вероятность появления существительных в подъязыке английской электроники равной 1/3 (см. § 1, п. 3), вычислим вероятности появления существительных в английском десятисловном сегменте ровно 0, 1, 2, ..., 10 раз.

Здесь $N = 10$, $p = f = 1/3$, $q = 2/3$. Используя только что приведенную схему, находим все $P'_N(x)$.

Ход вычисления иллюстрируется на примере расчета $P'(3)$.

1) Определим

$$Np = 10 \cdot \frac{1}{3} = \frac{10}{3};$$

$$\sigma = \sqrt{10 \cdot \frac{1}{3} \cdot \frac{2}{3}} = \frac{1}{3}\sqrt{20} = \frac{4{,}472}{3} = 1{,}491;$$

$$\frac{1}{\sigma} = \frac{1}{1{,}491} = 0{,}671.$$

2) Вычислим

$$= \frac{1}{\sqrt{Npq}}(x - Np) = 0{,}671 \cdot \left(3 - \frac{10}{3}\right) = -0.224.$$

3) Учитывая, что $\varphi(z)$ — четная функция, т. е. $\varphi(-z) = \varphi(z)$, по табл. II находим

$$\varphi(-0{,}224) = \varphi(0{,}224) = 0{,}3891.$$

4) Определяем произведение

$$\frac{1}{\sqrt{Npq}}\varphi(z) = 0{,}3891 \cdot 0{,}671 = 0{,}2611.$$

Остальные значения $P'_{10}(x)$, вычисленные по этой же схеме, приведены в столбце (3) табл. 6.2 (см. стр. 154). Сопоставляя эти вероятности с биномиальными вероятностями, нетрудно заметить, что асимптотическая формула (6.85) дает сравнительно небольшую погрешность даже при небольших значениях N.

Поскольку $Np = \mu$, а $\sqrt{Npq} = \sigma$, перепишем (6.85) в виде

$$P'_N(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} = f(x). \qquad (6.87)$$

Если равенство (6.87) рассматривать в качестве дифференциального закона (плотности вероятности) нормального распределения, то ему должен соответствовать, как было сказано в § 2, п. 5, некоторый интегральный закон (функция распределения). Таким интегральным законом нормального распределения является выражение

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x-\mu)^2/(2\sigma^2)}\, dx. \qquad (6.88)$$

Нормальное распределение имеет следующие параметры:

а) первый начальный момент (математическое ожидание)

$$\nu_1 = M(X) = \mu \qquad (6.89)$$

для ненормированной случайной величины и

$$\nu_1 = M(X) = M\left(\frac{x - Np}{\sqrt{Npq}}\right) = \frac{1}{\sigma} M(x - \mu) = 0 \qquad (6.90)$$

для нормированной случайной величины;

б) второй центральный момент (дисперсия)

$$\mu_2 = D(x) = \sigma^2 \qquad (6.91)$$

для ненормированной случайной величины и

$$\mu_2 = D\left(\frac{x - Np}{\sqrt{Npq}}\right) = 1$$

для нормированной случайной величины;

в) третий центральный момент, как и остальные нечетные моменты, для ненормированной и нормированной случайных величин равен нулю:

$$\mu_3 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x - \mu)^3 e^{-(x-\mu)^2/(2\sigma^2)}\, dx = 0, \qquad (6.92)$$

в связи с этим третий нормированный момент (коэффициент асимметрии) также равен нулю:

$$\rho_3 = \mu_3/\sigma^3 = 0; \qquad (6.93)$$

г) четвертый нормированный момент и эксцесс соответственно равны

$$\rho_4 = \mu_4/\sigma^4 = 3, \quad E = \mu_4/\sigma^4 - 3 = 0. \qquad (6.94)$$

Поскольку коэффициент асимметрии и эксцесс (крутизна) нормального распределения равны нулю, дифференциальная форма

нормального закона характеризуется идеальной по симметрии и крутизне кривой. Эта кривая имеет колоколообразную форму, расположена над осью абсцисс и симметрична относительно ординаты, отвечающей значению $x = \mu$. Если $\mu = 0$, то центром распределения служит начало координат.

При изменении величины μ и сохранении значения σ кривая, не меняя своей формы, передвигается вместе с точкой μ по оси Ox. При изменении величины σ и сохранении значения μ кривая, не меняя своего положения, меняет свою форму. При больших значениях σ кривая становится пологой, а при малых σ кривая, сжимаясь с боков, вытягивается вверх (рис. 46).

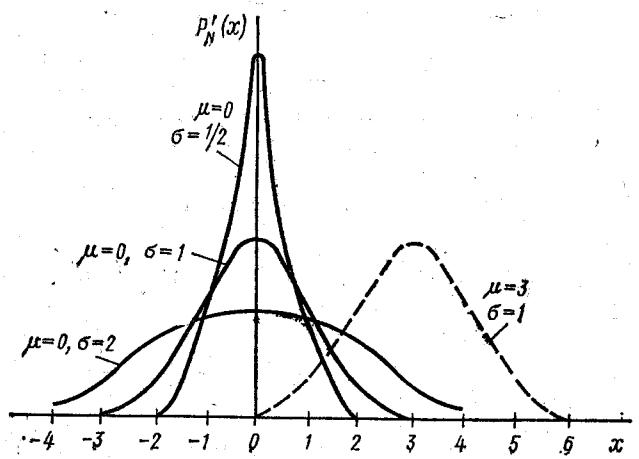


Рис. 46

Мы уже говорили (см. § 1, п. 7), что при решении лингвистических задач часто возникает необходимость определить вероятность того, что значение лингвистической случайной величины попадает в некоторый числовой промежуток от a до b, т. е. найти

$$P_N(B'_x) = P_N(a \leqslant X \leqslant b) =$$
$$= P_N(a) + P_N(a+1) + \ldots + P_N(b-1) + P_N(b).$$

Такая задача легко решается с помощью формулы Бернулли в тех случаях, когда N мало, а интервал от a до b не очень широк. Если же N велико, а интервал достаточно широк, использование указанной формулы приводит к чрезмерно громоздким вычислениям. В этих случаях используется приближенная асимптотическая формула

$$P(a \leqslant X \leqslant b) \approx \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-z^2/2} dz, \qquad (6.95)$$

получаемая из **интегральной предельной теоремы Муавра—Лапласа**, которая формулируется следующим образом.

Если вероятность появления события при каждом отдельном независимом испытании постоянна и равна p $(0<p<1)$ и число испытаний $N \to \infty$, то предел вероятности того, что число x появлений события заключено между $a = Np + z_1\sqrt{Npq}$ и $b = Np + z_2\sqrt{Npq}$, равен

$$\lim_{N\to\infty} P(a \leqslant X \leqslant b) = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-z^2/2} dz. \qquad (6.96)$$

Легко показать, что здесь

$$z_1 = \frac{a-Np}{\sqrt{Npq}} = \frac{a-\mu}{\sigma}, \quad z_2 = \frac{b-Np}{\sqrt{Npq}} = \frac{b-\mu}{\sigma}.$$

Решать конкретные лингвистические задачи с помощью формул (6.95) и (6.96) нельзя, поскольку интеграл $\int e^{-z^2/2} dz$ является неберущимся. Это затруднение преодолевается путем введения вспомогательного равенства

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-z^2/2} dz. \qquad (6.97)$$

Функция $\Phi(x)$ называется *функцией Лапласа*. В дальнейшем нам понадобятся следующие ее с в о й с т в а:

а) *нечетность*, т. е.

$$\Phi(-x) = -\Phi(x);$$

б) *монотонное возрастание*, т.е.

$$\Phi(x_1) < \Phi(x_2), \text{ если } x_1 < x_2;$$

в) *выполнение равенств*

$$\Phi(0) = 0,\ \Phi(-\infty) = -1/2,\ \Phi(+\infty) = 1/2.$$

Значения функции Лапласа приведены в табл. III, помещенной в Приложении (см. стр. 365).

Если воспользоваться теперь функцией Лапласа, то выражение (6.95) примет вид

$$P(a \leqslant X \leqslant b) \approx \frac{1}{\sqrt{2\pi}} \int_0^{(b-\mu)/\sigma} e^{-z^2/2} dz - \frac{1}{\sqrt{2\pi}} \int_0^{(a-\mu)/\sigma} e^{-z^2/2} dz =$$
$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \qquad (6.98)$$

При помощи только что описанной процедуры решим следующую лингвистическую задачу.

Опираясь на статистические данные [6, с.208], будем считать, что вероятность появления в русском тексте согласного звукотипа после согласного (комбинация СС) равна 0,26, в то время как появление гласного после согласного (группа СГ) равняется 0,74. Группы ГС

и ГГ при этом в расчет не принимаются. Определим вероятность того, что среди 500 случайным образом выбранных комбинаций СС и СГ окажется: а) от 120 до 150 звукокомбинаций СС; б) не менее 121 звукокомбинации СС; в) не более 150 звукокомбинаций СС; г) не более 119 звукокомбинаций СС; д) не менее 151 звукокомбинации СС.

Прежде всего заметим, что комбинации СС и СГ составляют полную группу событий, т. е. если СС принять за событие A, то СГ будет событием $\bar{A}$. Таким образом, по условию имеем:

$$P(A) = p = 0{,}26;\ P(\bar{A}) = q = 0{,}74;\ N = 500;\ a = 120;\ b = 150.$$

Найдем теперь пределы интегрирования:

$$z_1 = \frac{a - Np}{\sqrt{Npq}} = \frac{120 - 500 \cdot 0{,}26}{\sqrt{500 \cdot 0{,}26 \cdot 0{,}74}} = \frac{-10}{9{,}808} = -1{,}02,$$

$$z_2 = \frac{b - Np}{\sqrt{Npq}} = \frac{150 - 500 \cdot 0{,}26}{\sqrt{500 \cdot 0{,}26 \cdot 0{,}74}} = 2{,}04.$$

Из табл. III находим, что $\Phi(2{,}04) = 0{,}4793$, а $\Phi(-1{,}02) = -\Phi(1{,}02) = -0{,}3461$, откуда имеем

$$P(120 \leqslant X \leqslant 150) = 0{,}4793 + 0{,}3461 = 0{,}8254.$$

Аналогичным образом получаем ответы и к п. б) — д) данной задачи. В результате имеем:

$$P(X \geqslant 121) = 0{,}8212;\ P(X \leqslant 150) = 0{,}9793,$$
$$P(X \leqslant 119) = 0{,}1314;\ P(X \geqslant 151) = 0{,}0162.$$

Правильность полученных ответов предлагается проверить читателю.

5. Логнормальное распределение вероятностей длин текстовых словоупотреблений. Реальные распределения случайных лингвистических величин, характеризующиеся обычно правосторонней асимметрией, плохо аппроксимируются нормальным законом. В связи с этим делаются попытки моделировать эти эмпирические распределения с помощью распределений Кэпптейна, Шарлье [30], выравнивающих кривых Пирсона и Бородачева [32в, с. 335—360]. Такое моделирование должно опираться не на внешнее сходство опытной и теоретических кривых — его следует осуществлять, исходя из лингвистической сущности случайного явления или процесса, приводящего к тому или иному закону распределения. С этой точки зрения наибольший интерес представляет логарифмически-нормальное (логнормальное) распределение.

Основная идея лингвистического приложения логнормального распределения состоит в следующем.

Значение случайной лингвистической величины X обычно складывается не из независимых внутриязыковых и экстралингвистических величин, как это имеет место в случае нормального распределения. Чаще всего эти значения являются результатом действия ряда причин, производящих последовательные «импульсы», причем эффект этих импульсов зависит, с одной стороны, от интенсивности самих импульсов, а с другой — от величины X, созданной действием предыдущих импульсов. В этом случае нормально распределена не сама случайная величина X, а ее логарифм $\ln X$ [42, с.132—133].

Дифференциальный закон логнормального распределения имеет вид

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2/(2\sigma^2)}, \tag{6.99}$$

где величины μ и σ соответственно оцениваются с помощью равенств

$$\mu = \frac{\sum_{i=1}^{N} \ln x_i}{N} \tag{6.100}$$

и

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (\ln x_i - \mu)^2}{N-1}}. \tag{6.101}$$

Параметры этого распределения таковы:

а) первый начальный момент (математическое ожидание)

$$\nu_1 = \mu' = M(\ln X) = e^{\mu + \sigma^2/2}; \tag{6.102}$$

б) второй центральный момент (дисперсия)

$$\mu_2 = \sigma'^2 = D(\ln X) = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1); \tag{6.103}$$

в) нормированный момент третьего порядка (коэффициент асимметрии)

$$\rho_3 = A = \sqrt{e^{\sigma^2} - 1}\,(e^{\sigma^2} + 2); \tag{6.104}$$

г) четвертый нормированный момент и эксцесс соответственно равны

$$\rho_4 = (e^{\sigma^2} - 1)(e^{3\sigma^2} + 3e^{2\sigma^2} + 6e^{\sigma^2} + 6) + 3,$$
$$E = (e^{\sigma^2} - 1)(e^{3\sigma^2} + 3e^{2\sigma^2} + 6e^{\sigma^2} + 6). \tag{6.105}$$

Дифференциальная форма логнормального распределения характеризуется одновершинной (одномодальной) кривой и имеет правостороннюю (положительную) скошенность (рис. 47).

Теперь, опираясь на данные предварительного статистического эксперимента, попытаемся дать вероятностный прогноз появления в немецком тексте словоформ разной длины.

Наблюдения над текстами, а также данные интроспекции показывают, что хотя длина слова не имеет ясно выраженного предела (см. гл. 4, § 2), выбор слова определенной длины на данном шаге

текста зависит не только от семантических «импульсов» окружающего контекста, но также от длины предшествующего слова (ср. чередование в тексте длинных, коротких и средних по длине слов). Исходя из этих соображений, попробуем аппроксимировать распределение вероятностей длин словоупотреблений в немецком тексте с помощью логнормального закона.

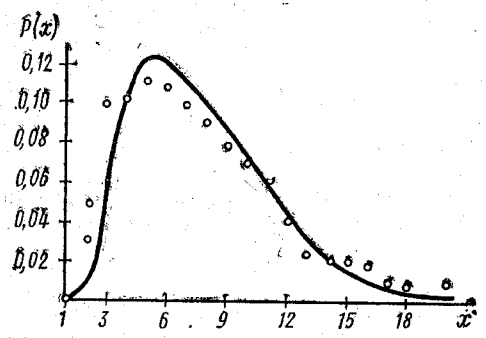


Рис. 47. Логнормальное распределение длин немецких словоформ:
———— теоретическая кривая;
○ ○ ○ ○ опытные данные

Предварительным экспериментом был охвачен научно-технический немецкий текст длиной в 1000 словоупотреблений; относительные частоты, полученные для словоформ длиной в x букв показаны в столбце (3) табл. 6.12. Опираясь на экспериментальные данные, с помощью выражений (6.100) и (6.101) получаем оценки $\mu = 2{,}3$ и $\sigma = 0{,}6$. Подставляя их в формулу (6.99), получаем вероятность появления двухбуквенных словоформ

$$P(2) = \frac{1}{2 \cdot 0{,}6\sqrt{2\pi}} e^{-(\ln 2 - 2{,}3)^2/(2 \cdot 0{,}36)} = \frac{1}{1{,}2 \cdot 2{,}51} e^{-(0{,}6931 - 2{,}3)^2/0{,}72} \approx$$
$$\approx 0{,}0091.$$

Остальные вероятности, полученные аналогичным путем приведены в столбце (2) табл. 6.12. Теоретическая кривая распределения длин немецких словоформ вместе с экспериментальными данными показана на рис. 47.

Таблица 6.12

Распределение вероятностей появления слов разной длины в немецком научно-техническом тексте

Длина слов в буквах (x)	$P(x)$	$f(x)$	Длина слов в буквах (x)	$P(x)$	$f(x)$
(1)	(2)	(3)	(1)	(2)	(3)
1	0,000	0,000	12	0,042	0,040
2	0,009	0,031	13	0,031	0,022
3	0,031	0,098	14	0,021	0,020
4	0,099	0,101	15	0,012	0,020
5	0,120	0,113	16	0,006	0,018
6	0,124	0,108	17	0,003	0,010
7	0,116	0,099	18	0,002	0,010
8	0,096	0,090	19	0,002	0,002
9	0,082	0,079	20	0,001	0,010
10	0,068	0,071		0,001	(и более)
11	0,054	0,060		...	

Только что полученное теоретическое распределение может быть использовано при построении вероятностных синтаксических алгоритмов и словаря немецко-русского машинного перевода.

Г. Хердан [51, с. 61—62] использует логнормальное распределение для математической экспликации вероятностного построения словаря языка и его реализации в тексте. По его мнению, эта логнормальность словаря и текста отражает присущий естественному языку принцип оптимального кодирования информации.

§ 4. Понятие о законе больших чисел

1. Принцип практической уверенности при лингвистическом исследовании. Сосредоточив внимание на построении вероятностных моделей текста, мы избегали вопроса о том, насколько рассмотренная нами модельная теория сходится с результатами лингво-статистического опыта. В частности, мы специально не касались вопроса о том, насколько вычисленная по приведенным выше моделям вероятность близка к полученной в опыте частости или в какой степени математическое ожидание соответствует средней частоте лингвистического события.

Однако каждая теория, использующая метод моделей, должна предусматривать процедуру, с помощью которой можно было бы оценить степень близости теоретических данных и экспериментальных результатов.

Выше уже отмечалось, что статистическое определение вероятности опирается на предположение о том, что если при N испытаниях событие A осуществлялось F раз, причем N очень велико, то отношение F/N должно быть близко к вероятности p события A (см. гл. 5, §3, п. 3). Это общее представление в какой-то степени связывает математическую теорию с лингво-статистической практикой, однако не дает возможности получить количественные оценки этой связи. Хотя мы и знаем, что частость может приближаться к вероятности, но пока не знаем, насколько именно. Задача состоит в том, чтобы неясное интуитивное предположение заменить математической процедурой, с помощью которой можно было бы количественно измерять отношение вероятностных данных к лингво-статистическим результатам.

Сформулировать такую процедуру можно, применяя понятия *практической достоверности и практической невозможности* события. В нашей повседневной деятельности, равно как и в лингвистическом поведении, мы постоянно опираемся на эти понятия. Садясь в самолет или в автомобиль, мы уверены в практической достоверности благополучного завершения нашего путешествия. Составляя русско-иноязычный математический словарь или формируя автоматический словарь для машинного перевода русских текстов по теории вероятностей, мы определенно не включим в русский словник существительное *дядя*, считая его появление в математических текстах практически невозможным. Вместе с тем благополучное окончание нашего воздушного путешествия или автомобильной прогулки

нельзя считать математически достоверным событием, поскольку единичные воздушные и автомобильные катастрофы пока имеют место. Аналогичным образом мы не гарантированы от того, что вопреки нормам русского подъязыка математики в некотором наугад взятом отрывке из математического сочинения не появится слово *дядя*; кстати, в этом можно убедиться, обратившись к 17-й снизу строке стр. 34 книги Е. С. Вентцель «Теории вероятностей» (изд. 3-е, М., 1964).

Несовпадение понятий математической достоверности (соответственно математической невозможности) события и практической ее достоверности (соответственно практической невозможности) лежит в основе несоответствия, существующего между логической структурой системы и вероятностным построением нормы языка. С точки зрения системы русского языка мы можем допустить, что в математических текстах возможно употребление словоформ *любовь, дядя, бубновый* и т. п. Вместе с тем, опираясь на норму, мы практически уверены, что эти формы не встретятся при чтении или машинном переводе математических текстов.

Понятия практической достоверности и невозможности, именуемые иногда *принципом практической уверенности*, лежат в основе всякого прогнозирования, в том числе прогнозирования, используемого в инженерной и прикладной лингвистике. Принцип практической уверенности можно сформулировать следующим образом: если вероятность события A в рассматриваемом процессе очень мала, то можно практически быть уверенным в том, что при однократном осуществлении опыта событие A не произойдет.

Вместе с тем следует помнить, что этот принцип не может быть задан в единой численной форме для всех областей деятельности человека. Дело в том, что пороги практической достоверности (невозможности) события определяются исходя из степени важности для нас того или иного события. Например, если при анализе и переводе иностранного текста лишь в шести предложениях из ста студент неправильно опознает синтаксическую структуру фразы, то можно быть практически уверенным, что, правильно пользуясь двуязычным словарем, он в целом поймет наугад взятый текст. Таким образом, вероятностью синтаксической ошибки, равной здесь 0,06, можно пренебречь. Что же касается перевода текста на ЭВМ, то такая же вероятность синтаксической ошибки не дает нам практической уверенности в правильности нашего автоматического перевода. Ведь машина не умеет, подобно человеку, пользоваться избыточностью текста и корректировать синтаксические ошибки, опираясь на лексический перевод. Верхняя граница вероятности допускаемой ошибки может быть приблизительно оценена здесь в 0,01 или 0,02. Еще ниже граница допустимой вероятности ошибки устанавливается в некоторых технических областях. Так, например, если бы вероятность аварийной посадки самолета была равна 0,06 и даже 0,01, то пассажирские перевозки на авиалиниях были бы, по-видимому, полностью исключены.

Итак, для каждой области применения теории вероятностей за-

даются свои, определяющиеся спецификой и практикой этой области принципы практической уверенности.

Одновременно вводится количественный критерий практической невозможности маловероятных событий, согласно которому предполагается, что случайное событие, вероятность которого ниже указанного порога, не произойдет при единичном испытании. Практическое использование этого критерия выглядит следующим образом. Предполагается, что вероятность события A находится ниже назначенного порога. Однако, произведя опыт, мы обнаруживаем, что событие A осуществилось. Тогда следует усомниться в справедливости исходной гипотезы и искать особую причину для появления события A. Например, появление в книге по теории вероятностей слова *дядя* заставляет нас усомниться в том, что мы имеем здесь место с отрывком, в котором рассматриваются чисто математические вопросы. И действительно, 17-я снизу строка на стр. 34 «Теории вероятностей» Е. С. Вентцель, где зафиксировано это слово, оказывается не математическим текстом, а цитатой из «Евгения Онегина»: «*Мой дядя самых честных правил...*».

2. Случайные и систематические ошибки в лингвистическом эксперименте. Отклонения опытных результатов от теоретических данных могут быть вызваны либо существенными факторами, вводимыми в структуру изучаемого лингвистического явления (примером может служить использование поэтической цитаты в научно-техническом тексте), либо ошибками измерения (так называемыми *систематическими ошибками*). И те и другие сравнительно легко обнаруживаются, и после их устранения результаты опыта, как правило, широко согласуются с данными теории.

Однако наряду с существенными отклонениями имеются такие отклонения значений случайной величины от ее математического ожидания, которые зависят от многих случайных, не учитываемых наблюдателем причин. К таким факторам, если говорить о коллективном восприятии речи, можно отнести личные стилистические вкусы говорящего, его настроение в момент произнесения или написания текста, обстановка, в которой происходит беседа, и т. п. Эти факторы, именуемые обычно *случайными ошибками*, то погашая, то усиливая друг друга, приводят к флуктуациям значений случайной величины относительно математического ожидания и соответственно вызывают колебания относительных частот вокруг вероятности.

В дальнейшем мы будем считать, что наши лингво-статистические измерения свободны от систематических ошибок: это позволит сосредоточить внимание на колебаниях частостей и частот, вызванных случайными ошибками.

Сама вероятность лингвистического события представляет для нас ценность лишь тогда, когда имеется практическая уверенность в том, что эта вероятность хорошо предсказывает и отражает результаты лингвистического эксперимента, или иначе говоря, что она очень близка к относительной частоте события, причем колебания частостей по отношению к вероятности невелики. Таким образом, основным вопросом соотношения теории вероятностей и лингво-ста-

тистической практики является вопрос о том, с какой вероятностью обеспечена сходимость относительной частоты к вероятности события.

Ответ на поставленный вопрос дается с помощью совокупности математических теорем, называемых законом больших чисел. Эти теоремы показывают связь между абстрактными моделями теории вероятностей и опытом, они дают возможность предсказывать результаты опытов. Наиболее общей формой закона больших чисел является теорема Чебышева.

3. Теорема Чебышева и другие формы закона больших чисел, их значение для лингвистического эксперимента. Прежде чем сформулировать основную теорему, рассмотрим некоторые предварительные понятия.

1) Пусть даны попарно независимые случайные величины X_1, X_2, ..., X_n, каждая из которых принимает свою последовательность частных значений: например, для X_1 имеем x_{11}, x_{12}, ..., x_{1l}, для X_2 имеем x_{21}, x_{22}, ..., x_{2l}, ..., для X_n имеем x_{n1}, x_{n2}, ..., x_{nl}. При этом распределение вероятностей значений для каждой из случайных величин может быть каким угодно. Найдем среднюю арифметическую этих величин:

$$\frac{X_1+X_2+\ldots+X_n}{n}=\overline{X}.$$

Легко показать, используя свойства математического ожидания (см. § 2, п. 8), что *математическое ожидание средней арифметической* $\overline{X}$ *случайных величин равно средней арифметической их математических ожиданий*:

$$M(\overline{X})=M\left(\frac{X_1+X_2+\ldots+X_n}{n}\right)=\frac{1}{n}M(X_1+X_2+\ldots+X_n)=$$
$$=\frac{1}{n}[M(X_1)+M(X_2)+\ldots+M(X_n)]. \qquad (6.106)$$

2) *Дисперсия средней арифметической* $\overline{X}$ *попарно независимых случайных величин равна средней арифметической этих дисперсий, уменьшенной в n раз*:

$$D(\overline{X})=D\left(\frac{X_1+X_2+\ldots+X_n}{n}\right)=\frac{1}{n^2}D(X_1+X_2+\ldots+X_n)=$$
$$=\frac{1}{n}\,\frac{D(X_1)+D(X_2)+\ldots+D(X_n)}{n}. \qquad (6.107)$$

Вывод выражения (6.107) опирается на свойства дисперсии (см. § 2, п. 8). По мере увеличения n значения средней арифметической $\overline{X}$, выступающей в роли случайной величины (ср. выше), будут все теснее группироваться вокруг математического ожидания $M(\overline{X})$. Уменьшение рассеивания здесь объясняется тем, что при образовании средней арифметической происходит частичное взаимное погашение случайных отклонений.

3. *Если дисперсия случайных величин ограничена некоторым положительным числом так, что* $D(X_1)\leqslant c$, $D(X_2)\leqslant c$, ..., $D(X_n)\leqslant c$, *то средняя арифметическая дисперсий не превосходит числа c:*

$$\frac{D(X_1)+D(X_2)+\ldots+D(X_n)}{n}\leqslant\frac{nc}{n}=c,$$

а дисперсия средней арифметической случайных величин не превышает величины $\frac{1}{n}$ *c:*

$$D\left(\frac{X_1+X_2+\ldots+X_n}{n}\right)\leqslant\frac{1}{n}c.$$

Используя только что введенные понятия, мы можем перейти к формулировке теоремы Чебышева, цель которой состоит в том, чтобы оценить отклонение средней арифметической случайных величин от математического ожидания, т. е. указать вероятность того, что это отклонение по абсолютной величине не будет превосходить заданного числа ε.

Теорема Чебышева. *Если* X_1, X_2, ..., X_n — *последовательность попарно независимых случайных величин, имеющих конечные дисперсии, ограниченные одной и той же постоянной c, т. е.* $D(X_1)\leqslant c$, $D(X_2)\leqslant c$, ..., $D(X_n)\leqslant c$, *то как бы мало ни было постоянное положительное число* ε, *с вероятностью, сколь угодно близкой к единице, можно утверждать, что отклонение средней арифметической этих n величин от средней арифметической их математических ожиданий не превосходит по абсолютной величине заданного положительного числа* ε, *если число n достаточно велико.*

Иными словами,

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}M(X_i)\right|\leqslant\varepsilon\right\}=1 \qquad (6.108)$$

и одновременно

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}M(X_i)\right|>\varepsilon\right\}=0. \qquad (6.109)$$

Теорема Чебышева имеет несколько важных для статистической и лингво-статистической практики частных случаев.

1. *Если все n независимых случайных величин имеют одинаковые математические ожидания, равные a, и их дисперсии ограничены одной и той же постоянной c, то как бы мало ни было положительное постоянное число* ε, *с вероятностью, сколь угодно близкой к единице, можно утверждать, что при* $n\to\infty$ *их средняя арифметическая стремится по вероятности к постоянной* $a=M(X_1)=M(X_2)=\ldots=M(X_n)$. Иными словами, имеет место неравенство

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i-a\right|\leqslant\varepsilon\right)\geqslant1-\frac{c}{n\varepsilon^2}, \qquad (6.110)$$

которое при неограниченном увеличении n превращается в равенство

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - a\right| \leqslant \varepsilon\right) = 1.$$

2. *Если вероятность наступления события A в каждом отдельном независимом испытании постоянна и равна p, то при достаточно большом числе испытаний (т. е. при $N \to \infty$) с вероятностью, сколь угодно близкой к единице, можно утверждать, что относительная частота F/N события A сколь угодно мало отличается от вероятности p,* т. е.

$$\lim_{N\to\infty} P\left(\left|\frac{F}{N} - p\right| \leqslant \varepsilon\right) = 1. \qquad (6.111)$$

Этот частный случай теоремы Чебышева известен под названием **теоремы Бернулли**.

Для доказательства этой теоремы используется вспомогательное неравенство, называемое **неравенством Чебышева:**

$$P(|X - M(X)| \geqslant \varepsilon) \leqslant \frac{D(X)}{\varepsilon^2}. \qquad (6.112)$$

Представим относительную частоту F/N в качестве случайной величины, математическим ожиданием которой служит вероятность p, а дисперсией — величина pq/N. Подставляя эти значения в неравенство Чебышева (6.112), после преобразований получаем

$$P\left(\left|\frac{F}{N} - p\right| \leqslant \varepsilon\right) \geqslant 1 - \frac{pq}{N\varepsilon^2}. \qquad (6.113)$$

Поскольку p, q, ε — заданные положительные числа и при $N \to \infty$ дробь $pq/(N\varepsilon^2)$ стремится к нулю, то приходим к соотношению (6.111). Пользуясь теоремой Бернулли, можно определять объемы выборок, необходимые для решения конкретных лингвистических задач.

Например, установлено [6, с. 238], что вероятность появления существительного в румынских текстах по радиоэлектронике равна 0,34, а допустимое абсолютное отклонение относительной частоты f от вероятности p равно 0,03. Определим тот наименьший объем исследуемого текста (наименьшую выборку), при котором заданные условия выполнялись бы с вероятностью 0,9545.

Здесь по условию $p = 0{,}34$; $\varepsilon = 0{,}03$; $P(|F/N - p| \leqslant \varepsilon) \geqslant 0{,}9545$; необходимо определить N. Подставляя эти данные в неравенство (6.113), имеем

$$0{,}9545 = 1 - \frac{0{,}34 \cdot 0{,}66}{N \cdot 0{,}03^2}, \text{ или } 1 - 0{,}9545 = \frac{0{,}34 \cdot 0{,}66}{N \cdot 0{,}0009},$$

откуда

$$N = \frac{0{,}34 \cdot 0{,}66}{0{,}0455 \cdot 0{,}0009} = 5473.$$

Таким образом, текст, необходимый для выполнения поставленных в задаче условий, должен содержать не менее 5473 словоупотреблений.

3. Теорема Бернулли характеризует соотношение между относительной частотой и постоянной вероятностью события. Однако часто мы имеем дело с такими лингвистическими явлениями, например со знаменательными словоформами и словосочетаниями, которые почти не повторяются в одних и тех же фиксированных условиях. Они многократно встречаются в тексте, но каждый раз в новых условиях разного лексического окружения, принципиально различных синтаксических позиций, причем вероятность интересующих нас лингвистических событий сильно зависит от этих меняющихся условий. Иными словами, мы имеем здесь дело с разными вероятностями $p_1, p_2, \ldots, p_N$ события A. Этот случай характеризует **теорема Пуассона.**

Если вероятность события A при каждом независимом испытании меняется, то при достаточно большом числе испытаний (т. е. при $N \to \infty$) с вероятностью, сколь угодно близкой к единице, можно утверждать, что относительная частота появления события A сколь угодно мало отличается от средней арифметической вероятностей $P = \frac{1}{N}\sum_{i=1}^{N} p_i$, лишь бы число испытаний было достаточно велико, т. е.

$$\lim_{N\to\infty} P\left(\left|\frac{F}{N} - \frac{1}{N}\sum_{i=1}^{N} p_i\right| \leqslant \varepsilon\right) = 1, \qquad (6.114)$$

где F — число появлений события A в N испытаниях.

В теоремах Бернулли, Пуассона и Чебышева закон больших чисел применяется к независимым величинам. Однако на практике (и особенно при статистическом описании текста) мы имеем дело с зависимыми или слабозависимыми величинами. В связи с этим возникает необходимость распространить закон больших чисел на зависимые величины. Эта задача решается с помощью **теоремы Маркова:** *если последовательность случайных зависимых или независимых величин, $X_1, X_2, \ldots, X_n$ такова, что при $n \to \infty$*

$$\frac{1}{n^2} D\left(\sum_{i=1}^{n} X_i\right) \to 0,$$

то при любом $\varepsilon > 0$ справедливо равенство

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} M(X_i)\right| \leqslant \varepsilon\right\} = 1. \qquad (6.115)$$

Доказательства всех приведенных теорем см. в работах [10, с. 287—297] и [14, с. 199—207].

На основании теоремы Бернулли или теоремы Пуассона иногда делается вывод, что с ростом числа испытаний частость неуклонно стремится к вероятности, т. е. что

$$\lim_{N\to\infty}\frac{F}{N}=p. \tag{6.116}$$

Такой вывод является совершенно необоснованным. Дело в том, что сходимость относительной частоты F/N к вероятности (ее обозначают термином «сходимость по вероятности») понимается здесь иначе, чем сходимость в математическом анализе (см. гл. 2, § 1, п. 2 и гл. 4, § 1, п.1). Различие между обоими понятиями сходимости заключается в следующем.

Если имеет место соотношение $\lim\limits_{N\to\infty}\frac{F}{N}=p$ в том смысле, который вкладывает в это равенство математический анализ, то это означает, что, начиная с некоторого $N=n$ и для всех последующих значений, выполняется неравенство

$$\left|\frac{F}{N}-p\right|=|f-p|\leqslant\varepsilon. \tag{6.117}$$

Когда же утверждается, что частость F/N сходится по вероятности (или стремится по вероятности) к p при $N\to\infty$, что записывается обычно так:

$$\frac{F}{N}\xrightarrow[N\to\infty]{\text{вер.}}p,$$

то для отдельных значений N неравенство (6.117) может и не выполняться. Дело в том, что в теореме Бернулли речь идет о вероятности одного единственного неравенства $\left|\frac{F}{N}-p\right|\leqslant\varepsilon$, которая при достаточно большом N становится больше, чем разность $1-\frac{pq}{N\varepsilon^2}$, или, иными словами, что при достаточно большом количестве испытаний частость будет как угодно мало отличаться от постоянной вероятности p.

Из сказанного видно, что теорема Чебышева, а также остальные ее частные случаи справедливы только для фиксированного значения $n=N$, например для $n_0=N_0$. В связи с этим нельзя утверждать, что неравенство

$$P\left(\left|\frac{F}{N_0}-p\right|\leqslant\varepsilon\right)\geqslant 1-\frac{pq}{N_0\varepsilon^2} \tag{6.118}$$

справедливо не только для N_0, но и для всех значений $N>N_0$. Однако для приложений теории вероятностей, и в частности для лингво-статистики, важно, чтобы теоремы закона больших чисел распространялись и на случаи $N>N_0$. Так, например, для лингво-статистики важно найти такие условия, накладываемые на неравенство (6.118), чтобы соотношение (6.116) было бы справедливо не только

относительно объема выборки N_0, но и для всех объемов выборки, где $N>N_0$. Чтобы достичь этого, на случайные величины X_i накладываются некоторые дополнительные условия.

Широкие условия для осуществления усиленного закона больших чисел определяются теоремами Колмогорова и Феллера [6, с. 244].

Первая из них показывает, что *достаточным условием для применения указанного закона к последовательности взаимно независимых случайных величин* $X_1, X_2, \ldots, X_N$ *является сходимость ряда* $\sum\limits_{N=1}^{\infty}\frac{D(X_N)}{N^2}$.

Вторая теорема говорит о том, что *усиленный закон больших чисел справедлив для последовательностей случайных величин, имеющих одинаковые распределения с конечным математическим ожиданием* $\mu=M(X_i)$.

Квантитативные исследования текста показывают, что численное значение, которое принимает та или иная случайная лингвистическая величина (частота или частость появления определенной фонемы, буквы, слога, словоформы, морфемы и т. д.), зависит от многих случайных причин, учесть которые мы не в состоянии. Если увеличить число случайных величин, то увеличится и число неучитываемых причин. Таким образом, становится, казалось бы, невозможным установить закономерности поведения суммы достаточно большого числа случайных лингвистических величин. Однако в действительности это не так: если индивидуальные величины содержат в себе более или менее значительный элемент случайности, то в их средней этот элемент взаимно погашается и становится исчезающе малым, если только количество усредняемых величин достаточно велико.

В итоге при некоторых сравнительно широких условиях суммарное поведение большого числа случайных лингвистических величин начинает утрачивать случайный характер и становится статистически устойчивым или, как говорят, почти закономерным. Статистическая закономерность начинает принимать здесь характер динамического закона.

Устойчивость средней арифметической, дающей при большом количестве испытаний сходимость по вероятности к математическому ожиданию, объясняет, почему и когда мы имеем право взять за истинное значение $M(X)$ значение средней арифметической. Вместе с тем становится ясным, почему при достаточном количестве испытаний можно с достаточной достоверностью использовать относительную частоту для оценки вероятности лингво-статистического события.

Одновременно закон больших чисел предостерегает лингвистов, пользующихся «симптоматической» статистикой, против неосмотрительного приравнивания частот и математических ожиданий лингвистических случайных величин, а также отождествления частостей и вероятностей независимо от того, имеются ли для этого условия, предусмотренные указанным законом.

На описанных свойствах средней арифметической и частости основан широко применяющийся в лингво-статистике (как, впрочем, и в других приложениях статистики) выборочный метод, сущность которого состоит в том, что по сравнительно небольшой случайной выборке текстов судят о целой разновидности языка — функциональной (стиль) или тематической (подъязык).

Сходимость средних арифметических частот, полученных по частичным выборкам, к математическим ожиданиям слов (или словосочетаниям) при достаточном числе выборок позволяет рассматривать частотные словари в качестве моделей вероятностного распределения слов и словосочетаний в норме данного подъязыка или стиля.

4. Центральная предельная теорема Ляпунова и сопоставление результатов лингвистического эксперимента с вероятностной речевой моделью. Теоремы, образующие первую часть закона больших чисел, давая полную степень практической уверенности о сходимости по вероятности определенных случайных величин к тем или иным постоянным, слишком завышают вероятность выполнения неравенства $|X - M(X)| > \varepsilon$. Одновременно занижается вероятность того, что отклонение случайной величины от ее математического ожидания будет не больше заданного порога ε. В связи с этим использование первой части закона больших чисел для нахождения таких характеристик, как точность, надежность оценки, доверительный интервал и т. д., связано с обследованием слишком больших текстовых выборок, объемы которых превосходят реальные возможности лингво-статистического исследования.

Поэтому возникает необходимость в такой процедуре, которая указывала бы более точно вероятности интересующих нас границ, используя при этом меньшее число испытаний, чем этого требуют теоремы закона больших чисел. Эта задача решается в центральной предельной теореме Ляпунова.

Центральная предельная теорема исходит из той же идеи, которая используется и при построении первой половины закона больших чисел. Эта идея заключается в том, что хотя исследуемое явление или процесс (в том числе и лингвистический) в ходе своей реализации подвергается действию большого числа независимых случайных воздействий, каждое из них лишь ничтожно мало изменяет ход процесса. Исследователь, интересующийся изучением процесса или явления в целом, а не воздействием отдельных факторов, должен наблюдать и фиксировать суммарное действие этих факторов.

В отличие от теоремы Чебышева, для которой характер распределения случайных величин $X_1, X_2, \ldots, X_N$, их сумм и средних $\bar{X} = (X_1 + X_2 + \ldots + X_N)/N$ не имеет значения, теорема Ляпунова утверждает, что каково бы ни было распределение независимых случайных величин, при определенных условиях распределение их средних подчиняется нормальному закону. Такой подход позволяет распространить на случаи, рассматриваемые в законе больших чисел, теорему Муавра—Лапласа, дающую возможность оценивать математические ожидания и вероятности появления отдельных значений случайной величины и таким образом более или менее точно определять вероятности отклонений $|\bar{X} - M(\bar{X})|$ и соответственно отклонений $\left|\frac{F}{N} - p\right| = |f - p|$.

Для того чтобы утверждение о нормальном распределении для средних имело место, достаточно, как показал А. М. Ляпунов, выполнение двух условий: во-первых, *все случайные слагаемые должны иметь конечные абсолютные центральные моменты третьего порядка*

$$M|X_i - M(X_i)|^3, \tag{6.119}$$

во-вторых, *отношение*

$$\frac{\sum_{i=1}^{N} M\,|X_i - M(X_i)|^3}{\left[\sum_{i=1}^{N} D(X_i)\right]^{3/2}} \tag{6.120}$$

должно стремиться к нулю при $N \to \infty$.

Смысл условий Ляпунова заключается в том, что ни одна из случайных величин, образующих среднюю, не была бы в ней преобладающей, во всяком случае, не была бы заметно больше других величин. Если же какая-либо величина или величины X_j, X_k оказывают преобладающее влияние на формирование $\bar{X}$, то второе условие не выполняется и утверждение о нормальном законе распределения средней не имеет места. Распределение средней здесь определяется законом распределения этих преобладающих случайных величин.

Лингвистическим примером этого явления может служить статистическое поведение так называемых *ключевых* (или *доминантных*) *слов* и словосочетаний текста, т. е. таких слов, которые передают основные понятия, рассматривающиеся в данном сообщении (в научно-технических текстах в качестве доминантных слов и словосочетаний выступают термины). Преобладающим фактором, влияющим на статистику доминантных единиц текста, является ситуация, отражаемая в содержании текста. Лингвистические, индивидуально стилевые и прочие факторы подавляются ситуацией. Так как появление тех или иных ситуаций не подчиняется нормальному закону*, то этому закону не подчиняются и распределения доминантных слов. Напротив, служебные слова, многие грамматические формы, фонемы и буквы, поведение которых определяется суммой большого числа случайных воздействий без преобладания в них семантики текста, дают, как правило, распределение, близкое к нормальному. Выделение в тексте слов, распределение которых не подчиняется нормальному закону, а также другим связанным с нормальным распределе-

* Во всяком случае, в рамках тех ограниченных «выборок ситуаций», которые представлены в текстах, написанных на естественном языке.

нием законам, лежит в основе эффекта статистического опознания терминологических единиц [32а, с.111]; см. также гл. 9, § 6, п.2.

После этих предварительных разъяснений перейдем к изложению центральной предельной теоремы Ляпунова и связанного с ней математического аппарата; ее доказательство см. в [14].

Теорема Ляпунова. *Пусть* $X_1, X_2, \ldots, X_N$ *— последовательность взаимно независимых случайных величин с конечными математическими ожиданиями* $M(X_1), M(X_2), \ldots, M(X_n)$ *и с конечными дисперсиями* $D(X_1), D(X_2), \ldots, D(X_n)$. *Тогда при выполнении условий* (6.119) *и* (6.120) *сумма этих случайных величин* $X^* = \sum_{i=1}^{n} X_i$ *с достаточной степенью точности распределена по нормальному закону с параметрами*

$$M(X^*) = \mu = M(X_1) + M(X_2) + \ldots + M(X_n)$$

и

$$D(X^*) = \sigma^2 = D(X_1) + \ldots + D(X_n).$$

Отсюда вероятность того, что случайная величина X^* примет какое-либо значение в промежутке (x_1, x_2), согласно формуле (6.98) составляет

$$P(x_1 < X^* < x_2) = \frac{1}{\sqrt{2\pi}} \int_{(x_1-\mu)/\sigma}^{(x_2-\mu)/\sigma} e^{-z^2/2}\, dz = \Phi\left(\frac{x_2-\mu}{\sigma}\right) - \Phi\left(\frac{x_1-\mu}{\sigma}\right). \quad (6.121)$$

Частным случаем последнего выражения служит равенство

$$P(|X^* - \mu| < z\sigma) = 2\Phi(z). \quad (6.122)$$

Нетрудно догадаться, что теорема Ляпунова имеет место и тогда, когда случайная величина является суммой достаточно большого числа одинаково распределенных независимых случайных величин, имеющих абсолютные центральные моменты третьего порядка. При доказательстве этого утверждения следует учитывать тот факт, что моменты всех порядков этих случайных величин — в том числе центральные моменты второго и третьего порядков — совпадают.

Средняя арифметическая указанных величин с учетом того, что $n = N$, равна

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i,$$

ее дисперсия

$$D(\overline{X}) = \sigma_{\overline{X}}^2 = D\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2}\sum_{i=1}^{N} D(X_i) = \frac{1}{N^2}(\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_N^2) = \frac{\sigma^2}{N}, \quad (6.123)$$

а среднее квадратическое отклонение

$$\sigma_{\overline{X}} = \sigma/\sqrt{N}. \quad (6.124)$$

Используя в соотношении (6.122) вместо σ только что полученное значение $\sigma_{\overline{X}}$, а вместо X^* — среднюю арифметическую $\overline{X}$, приходим к равенству

$$P\left(|\overline{X} - \mu| < z\frac{\sigma}{\sqrt{N}}\right) = 2\Phi(z),$$

которое можно переписать в виде

$$P\left(\left|\frac{\overline{X}-\mu}{\sigma}\sqrt{N}\right| < z\right) = 2\Phi(z). \quad (6.125)$$

Примерами таких случайных величин могут являться случайные лингвистические величины, распределенные по простой схеме повторений испытаний. Следовательно, здесь в качестве усредненной случайной величины можно рассматривать относительную частоту $F/N = f$, математическое ожидание которой есть вероятность p. При этом нормированное отклонение примет вид

$$\frac{f-p}{\sigma} = \frac{f-p}{\sqrt{pq/N}} \quad (6.126)$$

(ср. с § 3, п.4).

Если N достаточно велико, то, учитывая (6.124), можно переписать выражение (6.122) в виде

$$P\left(\left|\frac{f-p}{\sqrt{pq/N}}\right| < z\right) = 2\Phi(z), \quad (6.127)$$

где

$$z = \frac{\varepsilon}{\sqrt{pq/N}} = \varepsilon\sqrt{\frac{N}{pq}}. \quad (6.128)$$

Отсюда

$$\varepsilon = z\sqrt{\frac{pq}{N}} = z\sqrt{\frac{p(1-p)}{N}} \quad (6.129)$$

и

$$N = z^2 pq/\varepsilon^2. \quad (6.130)$$

При решении лингвистических задач, оперирующих нормально распределенными случайными величинами X, $\overline{X}$ или X^*, целесообразно задать такой интервал рассеяния случайной величины вокруг ее математического ожидания (соответственно частостей вокруг вероятности), в который попадало бы большинство значений случайной величины.

С этой целью, записав равенство (6.122) в виде

$$P(|X - \mu| < z\sigma) = 2\Phi(z),$$

придадим z целочисленные значения 1, 2, 3. Тогда, учитывая данные табл. III (см. стр. 365), получим

при $z = 1 : P(|X - \mu| < \sigma) = 2\Phi(1) = 2 \cdot 0{,}3413 = 0{,}6826;$ (6.131)

при $z = 2 : P(|X - \mu| < 2\sigma) = 2\Phi(2) = 2 \cdot 0{,}4772 = 0{,}9544;$ (6.132)

при $z = 3 : P(|X - \mu| < 3\sigma) = 2\Phi(3) = 2 \cdot 0{,}49865 = 0{,}9973.$ (6.133)

В лингвистических задачах обычно используется выражение (6.133), известное под названием «*правила трех сигм*». Это правило утверждает с вероятностью $2\Phi(3) = 0{,}9973$, что отклонения случайной величины X от ее математического ожидания не превосходят величины 3σ. Вероятность же того, что отклонение случайной величины X выйдет за пределы трехкратного среднеквадратического отклонения, равно 0,0027. Иными словами, здесь имеется практическая уверенность, что погрешности лингвистического наблюдения не превысят заданной ошибки наблюдения $\varepsilon = 3\sigma$.

В некоторых работах используются более узкие — двухсигмовый (6.132) и односигмовый (6.131) интервалы [34]. В этих случаях практическая уверенность в том, что случайная величина попадет в заданный интервал, значительно меньше.

До сих пор мы рассматривали вопрос о применимости центральной предельной теоремы к независимым величинам. Можно доказать [6, с. 251], что центральная предельная теорема может быть распространена и на зависимые случайные величины при условии, что связь между величинами $X_1, X_2, \ldots, X_i, \ldots, X_k, \ldots, X_N$ постепенно ослабевает по мере удаления их друг от друга, т. е. при возрастании разности $i - k$. Этот результат представляет прямой интерес для лингвистического исследования речи, имеющей дело со слабо зависимыми величинами. В целом же на центральной предельной теореме Ляпунова и ее следствиях основываются как выборочный метод, так и сопоставление результатов лингвистического эксперимента с вероятностными моделями построения текста.

Примерам этого сопоставления будут посвящены следующие главы.

ГЛАВА 7

ПЕРВИЧНАЯ СТАТИСТИЧЕСКАЯ ОБРАБОТКА ТЕКСТА

§ 1. Статистическая совокупность лингвистических объектов и ее организация

Исследование текста с помощью описанных выше вероятностных моделей может быть осуществлено при том условии, что произведена первичная статистическая обработка текста и к ее результатам применены специальные критерии перехода к вероятностной модели.

Прежде всего познакомимся с приемами первичной статистической обработки текстов.

1. Статистическая совокупность лингвистических объектов. Всякое статистическое исследование предусматривает наблюдение над множеством объектов. Эти объекты характеризуются многими признаками, каждый из которых варьируется при переходе от одного объекта к другому. Все признаки одновременно рассмотреть невозможно, поэтому языковед должен сосредоточить свое внимание на одном определенном признаке, предполагая, что в отношении остальных признаков объекты данного лингвистического множества равноправны. Используя такое допущение, мы можем считать, что рассматриваемое множество однородно. Построенное указанным способом множество называется *статистической совокупностью*, а составляющие ее объекты — *единицами совокупности*.

Лингвистические объекты обладают как количественными, так и качественными свойствами. Количественные свойства (например, длина словоформы в буквах или фонемах, слогах, морфемах, либо количество словоупотреблений в предложении и т. п.) постоянно используются в качестве тех признаков, по которым лингвистические объекты выступают в качестве единиц статистической совокупности.

Однако статистика текста оперирует не только количественными, но и качественными признаками. Например, в ходе статистико-морфологического исследования словоупотребления текста группируются по признаку их принадлежности к той или иной части речи. При статистико-синтаксическом исследовании таким качественным признаком является функционирование каждого словоупотребления в роли определенного члена предложения.

Часто бывает удобно использовать лишь два качественных признака, точнее признак A и его отсутствие — *не* A ($\bar{A}$). В этом случае говорят об *альтернативном качественном признаке*. В только что рассмотренном примере в качестве альтернативного признака можно рассматривать отнесение данного словоупотребления к существительным или не-существительным (соответственно к глаголу или не-глаголу, подлежащему или не-подлежащему и т. п.).

Отдельные лингвистические статистические совокупности могут образовывать вместе более крупную совокупность — совокупность совокупностей. Одновременно каждая совокупность может состоять

из частных совокупностей, которые в свою очередь могут рассматриваться как единицы совокупности.

Если статистическая совокупность объединяет все однородные лингвистические объекты, обладающие данным качественным или количественным признаком (признаками), то такую совокупность называют *генеральной лингвистической совокупностью*. Генеральная совокупность может содержать как конечное, так и бесконечное количество единиц. Если генеральная совокупность бесконечна или очень велика, то исследованию подвергается некоторая обозримая ее часть, называемая *выборочной лингвистической совокупностью* (*выборкой*).

Например, если признаком объекта является длина словоформы в пушкинском тексте, то в качестве генеральной совокупности выступают все тексты, написанные А. С. Пушкиным. Отдельные же произведения, например «Капитанская дочка», являются выборками, извлеченными из генеральной совокупности. Если же исследуется распределение длин словоформ в русском литературном языке, то генеральной совокупностью служит сумма всех русских литературных текстов. Заметим, что если хронологические границы существования русского литературного языка не фиксированы, то число словоупотреблений (однородных объектов) здесь бесконечно. Произведения А. С. Пушкина в этом случае выступают в виде конечной выборочной лингвистической совокупности внутри бесконечной генеральной совокупности языковых объектов.

2. Методы организации статистического наблюдения над текстом. Успех каждого лингвистического исследования зависит от организации статистического наблюдения, которая предусматривает, во-первых, выбор лингвистического признака и установление единицы совокупности, во-вторых, определение способа наблюдения.

Само собой разумеется, что каждый количественный или качественный признак, применяемый для выделения единицы совокупности, должен иметь лингвистический смысл и отвечать задачам данного языковедческого исследования. Часто случается, что в лингвостатистике используются такие качественные критерии, при которых граница переходов от одного состояния к другому оказывается весьма неопределенной, например деление слов текста на знаменательные (полнозначные) и служебные.

Однако каким бы ни было основание для группировки — естественное или искусственное, определенное или неопределенное — конечное решение должно быть всегда строго фиксированным. Каждый лингвистический объект должен быть признан либо обладающим, либо не обладающим данным качественным признаком.

Статистическое наблюдение предусматривает с п л о ш н о е и в ы б о р о ч н о е обследование генеральной совокупности. Сплошное обследование используется в лингво-статистике тогда, когда, во-первых, генеральная совокупность хотя и велика, но все же обозрима, во-вторых, когда необходимо учесть все употребления интересующих нас языковых объектов, например слов. Такая ситуация имеет место при статистическом описании языка писателя (частотные словари произведений А. С. Пушкина или трудов основоположника современной казахской литературы Абая Кунанбаева) и при исследовании языка отдельного художественного произведения (частотные словари романа Дж. Джойса «Уллис» или «Стихов о прекрасной даме» А. Блока) — см. [7, с.10].

Обычно же генеральная совокупность настолько велика, что применить сплошное обследование оказывается невозможным даже при условии использования вычислительной техники. Поэтому здесь применяется лишь часть единиц генеральной совокупности. Такое наблюдение может быть осуществлено с помощью либо повторной, либо бесповторной выборки (см. гл. 6, § 1, п.1). И в том , и в другом случае имеется в виду перенос результатов наблюдения над частотной выборкой на всю генеральную совокупность. Этот перенос может быть осуществлен в том случае, если средняя величина признака и его относительная частота (доля) в выборочном наблюдении достаточно хорошо воспроизводит среднюю величину и долю признака в генеральной совокупности.

Статистика предполагает следующие приемы выборочного наблюдения.

1. С л у ч а й н ы й о т б о р. Здесь выбор отдельных единиц осуществляется либо по жребию, путем подбрасывания монет или игральной кости и т. д., либо путем использования таблиц случайных чисел. При этом каждая единица совокупности имеет равную возможность попасть в выборку. Это обеспечивает достаточную близость средней выборочной величины к средней генеральной величине. Этот вид отбора ввиду его громоздкости сравнительно редко используется в лингвистике.

2. М е х а н и ч е с к и й о т б о р. Здесь единицы совокупности выбираются в определенном, формально установленном порядке. Например, желая исследовать распределение гласных, мы нумеруем все фонемы текста, после чего фиксируем присутствие или отсутствие гласной во всех фонемных позициях, номер которых кратен 10 (или 5, 3 и т. п.).

3. С е р и й н ы й о т б о р. В противоположность рассмотренным выше видам выборки, где отбор каждой единицы проводится в индивидуальном порядке, серийная выборка предполагает отбор сериями. Эти серии отбираются в случайном порядке, чаще бесповторным способом. Отобрав таким образом серии, исследователь проводит внутри их сплошное наблюдание (см. гл. 6, § 1).

4. Т и п и ч е с к и й о т б о р. Общий недостаток первых трех приемов выборочного обследования текста состоит в том, что они не учитывают смысловых и жанрово-стилистических своеобразий отдельных частей текста, выступающего в роли генеральной совокупности. Эти различия оказывают заметное воздействие на статистику знаков — в первую очередь слов, словоформ и словосочетаний. Так, например, относительные частоты существительного *крепость*, местоимения *я* и первого лица глаголов в «Капитанской дочке» значительно выше, чем во всем пушкинском тексте. Типический отбор предполагает предварительную разбивку генеральной совокуп-

ности по определенному признаку на однородные тематические группы, из которых затем случайным порядком выбираются интересующие нас лексические или грамматические единицы. При сопоставлении частотных словарей типическая выборка сочетается с серийным отбором. Количество серий, извлекаемых из каждой тематической группы, определяется удельным весом этой группы в генеральной совокупности.

§ 2. Вариационные ряды лингвистических признаков

1. Дискретные вариационные ряды. В ходе наблюдения мы получаем сведения о количественном или качественном изменении изучаемого признака относительно каждой единицы нашей совокупности. Так, например, для определения длины словоформы в казахских публицистических текстах взято подряд сто словоупотреблений из передовой статьи газеты «Қазахстан мугалими» (26.X. 1969 г.). В результате получена следующая последовательность чисел, каждое из которых характеризует длину словоупотребления в буквах:

3,	6,	4,	7,	10,	13,	6,	8,	4,	4,
6,	9,	10,	10,	7,	6,	5,	9,	11,	9,
5,	4,	8,	8,	3,	7,	8,	3,	11,	11,
7,	9,	5,	12,	6,	11,	8,	8,	7,	8,
11,	5,	6,	5,	5,	7,	8,	8,	8,	7,
5,	7,	6,	6,	6,	5,	9,	3,	11,	11,
16,	7,	11,	11,	3,	3,	9,	9,	10,	3,
5,	13,	12,	6,	8,	6,	6,	3,	7,	4,
9,	3,	12,	11,	6,	14,	6,	10,	16,	8,
9,	8,	7,	9,	4,	5,	3,	10,	8,	3.

Порядок следования чисел повторяет здесь последовательность словоупотреблений в тексте. Рассматривая приведенную в примере последовательность чисел, нетрудно заметить, что величина интересующего нас признака (длины словоформ) варьирует от одной единицы совокупности (словоупотребления) к другой. Задачей статистического наблюдения, в том числе и лингво-статистического, является изучение вариации признака (варьирующего признака) в данной совокупности.

Вернемся к рассмотрению примера. Как уже говорилось, роль варьирующего признака выполняет длина словоупотребления, причем для каждого из ста словоупотреблений этот признак принимает свое значение (3, 6, 4 и т. д. букв). Возможные значения признака в статистике называются *вариантами*. Различия между вариантами могут быть как количественными (дискретными или непрерывными), так и качественными (см. ниже).

Если ранжировать варианты нашего признака, расположив их по возрастанию, то получим такую последовательность длин словоупотреблений:

3,	3,	3,	3,	3,	3,	3,	3,	3,	3,
3,	4,	4,	4,	4,	4,	4,	5,	5,	5,
5,	5,	5,	5,	5,	5,	5,	6,	6,	6,
6,	6,	6,	6,	6,	6,	6,	6,	6,	6,
6,	7,	7,	7,	7,	7,	7,	7,	7,	7,
7,	7,	8,	8,	8,	8,	8,	8,	8,	8,
8,	8,	8,	8,	8,	8,	9,	9,	9,	9,
9,	9,	9,	9,	9,	9,	10,	10,	10,	10,
10,	10,	11,	11,	11,	11,	11,	11,	11,	11,
11,	11,	12,	12,	12,	13,	13,	14,	16,	16.

Ранжирование может быть осуществлено не только по возрастанию, но и по убыванию значений признака.

Ранжированная запись всегда слишком длинна и громоздка. Компактнее и нагляднее представить варьирование признака в виде таблицы, в верхней строке которой указываются значения признака (варианты), а в нижней — число повторений данного значения. Полученная в результате такого вторичного упорядочения таблица называется *вариационным рядом* (*рядом распределений* или *эмпирическим распределением признака*).

Вариационный ряд длины казахских словоформ по тексту из газеты «Қазахстан мугалими» показан в табл. 7.1.

Таблица 7.1

Длина словоформы	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Всего
Число повторений словоформы данной длины	11	6	10	14	11	14	10	6	10	3	2	1	0	2	100

Обычно признак обозначается большими буквами латинского алфавита X, Y, ..., а варианты — соответствующими строчными буквами $x_1, x_2, \ldots, x_k, y_1, y_2, \ldots, y_k, \ldots$; число повторений вариант — через $n_1, n_2, \ldots, n_k$. Сумма всех вариант N равна в этом случае

$$N = n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i.$$

Общий вид вариационного ряда показан в табл. 7.2.

Таблица 7.2

x_1	x_2	...	x_k	Всего
n_1	n_2	...	n_k	N

Вариационные ряды, представленные в табл. 7.1 и 7.2, содержат в нижней строке абсолютные частоты n_i, однако вместо абсолютных частот можно указывать относительные частоты (частости) $f_i = \frac{n_i}{N}$ или даже проценты $f_i \cdot 100\%$ (табл. 7.3).

Таблица 7.3

x_1	x_2	...	x_k	Всего
f_1	f_2	...	f_k	1
$f_1 \cdot 100\%$	$f_2 \cdot 100\%$	...	$f_k \cdot 100\%$	100%

В только что рассмотренном примере мы имели сравнительно небольшое варьирование признака (всего 14 вариант). Однако часто приходится иметь дело с несколькими десятками и даже сотнями вариант. В этом случае вариационный ряд получается очень растянутым и поэтому плохо обозримым. Чтобы избежать этого неудобства, в верхней строке таблицы указываются не сами значения признака, а интервалы, в которых находятся эти значения. В нижней строке указывается, сколько вариант падает на один интервал.

Например, при подсчете распределения существительных в 500 случайных выборках по 100 словоупотреблений каждая, взятых из немецких текстов по физической химии, получен слишком длинный вариационный ряд (см. табл. 7.4).

Таблица 7.4

X (количество существительных в одной сотне словоупотреблений)	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
N (количество сотен)	1	0	2	5	6	10	15	20	26	23	31	41	33	35	51

Продолжение табл. 7.4

X (количество существительных в одной сотне словоупотреблений)	28	29	30	31	32	33	34	35	36	37	38	39	40	41	Всего
N (количество сотен)	34	33	33	24	19	16	14	12	7	2	4	0	1	2	500

Таблица 7.5

Интервалы вариант	12—13	14—15	16—17	18—19	20—21	22—23	24—25	26—27	28—29	30—31	32—33	34—35	36—37	38—39	40—41	Всего
Частоты n_i	1	2	11	25	46	54	74	86	67	57	35	26	9	4	3	500

Таблица 7.6

Интервалы вариант	12—14	15—17	18—20	21—23	24—26	27—29	30—32	33—35	36—38	39—41	Всего
Частоты n_i	1	13	45	80	109	118	76	42	13	3	500

Чтобы сделать этот ряд более компактным и обозримым, берем интервалы, содержащие по два (табл. 7.5) и по три (табл. 7.6) значения признака. Наиболее наглядную картину распределения дает, очевидно, вторая таблица.

2. Непрерывные вариационные ряды. В лингвистике используются не только дискретные, но и непрерывные вариационные ряды. Эти последние используются, как правило, при исследовании фонетических явлений, поскольку здесь значения признака (например длина, частота, интенсивность звука) могут отличаться друг от друга на как угодно малую (бесконечно малую) величину. Поскольку различия между вариантами имеют непрерывный характер, в этом случае используется только интервальное построение вариационного ряда.

Рассмотрим следующий пример. В ходе исследования длины китайского слога произведено 150 замеров времени звучания слогов, произнесенных дикторами-китайцами, причем длины слогов колеблются от 40 до 300 мс [7, с. 22]. В итоге вторичного упорядочения этих данных получено три вариационных ряда (см. табл. 7.7—7.9).

Таблица 7.7

Интервалы вариант (длины слогов в мс)	40—50	50—60	60—70	70—80	80—90	90—100	100—110	110—120	120—130	130—140	140—150	150—160	160—170	170—180
Частоты n_i	2	0	5	2	4	3	16	13	20	9	17	7	11	6

Продолжение табл. 7.7

Интервалы вариант (длины слогов в мс)	180—190	190—200	200—210	210—220	220—230	230—240	240—250	250—260	260—270	270—280	280—290	290—300	300—310	Всего
Частоты n_i	5	6	10	2	6	1	0	1	2	0	1	0	1	150

Таблица 7.8

Интервалы вариант (длины слогов в мс)	40—60	60—80	80—100	100—120	120—140	140—160	160—180	180—200	200—220	220—240	240—260	260—280	280—300	300—320	Всего
Частоты n_i	2	7	7	29	29	24	17	11	12	7	1	2	1	1	150

Таблица 7.9

Интервалы вариант (длины слогов в мс)	40—70	70—100	100—130	130—160	160—190	190—220	220—250	250—280	280—310	Всего
Частоты n_i	7	9	49	33	22	18	7	3	2	150
Накопленные частоты n_i^*	7	16	65	98	120	138	145	148	150	

В общем случае непрерывный интервальный ряд задается так, как это показано в табл. 7.10.

Таблица 7.10

Значения признака X	(x_1, x_2)	(x_2, x_3)	...	(x_m, x_{m+1})
Частоты	n_1	n_2	...	n_m
Частости	f_1	f_2	...	f_m

Здесь (x_1, x_2), (x_2, x_3), ..., (x_m, x_{m+1}) являются интервалами, на которые разбиты возможные значения признака. Введем понятие *интервальных разностей* $k_1 = x_2 - x_1$, $k_2 = x_3 - x_2$, ..., $k_m = x_{m+1} - x_m$, характеризующих ширину интервалов. Если интервалы имеют одинаковую ширину, то интервальные разности равны, т. е. $k_1 = k_2 = \ldots = k_m$.

В тех случаях, когда вариационный ряд имеет разные по величине интервалы, целесообразно пользоваться понятием *плотности распределения*, представляющей собой частоту, приходящуюся на единицу величины интервала:

$$v_i = n_i/k_i.$$

Вместо плотности интервала можно пользоваться также *относительной плотностью распределения*

$$\varphi_i = f_i/k_i.$$

Желательно, чтобы длина интервала, а также его границы и середины были выражены целым или округленным числом. Так как все интервалы должны иметь одинаковую длину, то совершенно очевидно, что начало первого интервала не обязательно должно совпадать со значением минимальной варианты — таков, например, вариационный ряд, описывающий распределение существительных в немецких технических текстах. Аналогичным образом, конец последнего интервала может и не совпадать со значением максимальной варианты, как это имеет, например, место в вариационных рядах, описывающих распределение длин китайских слогов. Вообще же при построении интервального ряда необходимо определять интервал настолько ясно, чтобы не оставалось никаких сомнений в отношении его границ и не мог бы возникнуть вопрос о том, к какой группе принадлежит та или другая варианта.

Не менее важным является вопрос об определении величины интервала при выборе количества интервалов. При уменьшении количества интервалов (соответственно при увеличении ширины интервала) общий вид распределения становится более наглядным (см. табл. 7.6—7.9), однако теряется информация о промежуточных вариациях признака внутри большого интервала. Эту информацию можно получить, сужая интервалы и увеличивая тем самым их число, но при этом таблица становится громоздкой и труднообозримой (см. табл. 7.7).

Выбор ширины интервала и их числа должен осуществляться таким образом, чтобы выделить характерные особенности распределения и сгладить случайные колебания. При решении этой задачи в лингво-статистике используются два приема.

Во-первых, ширина интервала может быть определена с помощью *формулы Стерджесса* [7, с. 24]

$$k = \frac{x_{\max} - x_{\min}}{1 + \log_2 N}. \qquad (7.1)$$

При этом интервальная разность k округляется до ближайшего целого числа. Число же интервалов определяется из выражения

$$l = \frac{x_{\max} - x_{\min}}{k} = 1 + \log_2 N. \qquad (7.2)$$

Если же минимальная и максимальная варианты оказываются за пределами полученных интервалов, то соответственно добавляются два интервала — один слева, другой справа.

Во-вторых, при определении числа и ширины интервалов можно пользоваться эмпирическими соответствиями, приведенными в табл. 7.11.

Используя формулы (7.1) и (7.2), а также данные табл. 7.11, проверим корректность нашего интуитивного построения вариационного ряда немецких существительных (см. табл. 7.5 и 7.6).

Таблица 7.11

Количество вариант	Число интервалов	Количество вариант	Число интервалов
25— 40	5— 6	100—200	8—12
40— 60	6— 8	более 200	10—15
60—100	7—10		

Подставляя соответствующие значения в формулу (7.1), имеем

$$k = \frac{41-13}{1+\log_2 500} = \frac{28}{1+8{,}96} \approx 2{,}81;$$

если округлить полученный результат до трех, оказывается, что он соответствует выбранной в табл. 7.6 ширине интервала.

Число интервалов $l = 1 + \log_2 500 \approx 10$ также соответствует количеству групп в табл. 7.6, которое оказывается, однако, заниженным по сравнению с рекомендацией табл. 7.11.

3. Порядковый признак у лингвистических единиц. В лингвистических исследованиях часто встречаются такие ситуации, когда дать точную количественную характеристику признака либо невозможно, либо нецелесообразно. В то же время условия эксперимента позволяют нам ранжировать варианты, т. е. расположить их в определенном порядке.

Например, при проведении эксперимента по индивидуальному угадыванию текста испытуемый не может назвать вероятности появления букв в той или иной позиции текста. Однако языковое чутье позволяет ему довольно точно указывать, какая буква является наиболее вероятной в данной позиции, а какие буквы по вероятности их появления стоят на втором, третьем и т. д. местах [23, с.12—15, 44—47, 53—55].

Аналогичным образом при составлении частотных словарей, опирающихся на малые выборки, пользоваться абсолютными и относительными частотами отдельных слов и словосочетаний нецелесообразно, поскольку статистическая ошибка при определении этих частот слишком велика. В этих случаях рассматривается порядок (ранг) расположения отдельных словоформ или словосочетаний.

Ранжирование широко используется в лингво-психологических исследованиях. В частности, этот прием применяется при коллективном тестировании, причем в итоге выводятся «коллективный» ранг для вариант исследуемого признака.

Этот прием использован Р. М. Фрумкиной [7, с.27—28] при сравнении объективных (статистических) и субъективных (интуитивных) оценок вероятностей слов. В качестве экспериментального материала были взяты десять слов, ранжированных по убыванию их частот согласно данным «Частотного словаря современного русского языка» Э. А. Штейнфельдт [39]. Этим путем получено объективное (статистическое) ранжирование. Для определения субъективного

ранжирования указанные слова были переданы десяти преподавателям русского языка, каждый из которых должен был, опираясь на свою лингвистическую интуицию, ранжировать эти слова по убыванию их вероятности. Индивидуальные ранги суммировались по отдельным словам. Каждая сумма рассматривалась как число баллов, количественно характеризующее соответствующее слово. Затем производилось вторичное ранжирование по возрастанию количества баллов. Слову, набравшему наименьшую сумму баллов, приписывался ранг 1; слову, имеющему следующую по величине сумму, был дан ранг 2 и т. д. Ход обработки результатов эксперимента показан в табл. 7.12.

Таблица 7.12

Ранговое сравнение субъективных и статистических оценок вероятностей слов

Слова	Испытуемые										Сумма рангов, предложенных испытуемым	Ранг по результатам эксперимента	Ранг по данным словаря Штейнфельдт
	1	2	3	4	5	6	7	8	9	10			
Сказать	4	2	1	3	2	3	3	1	4	1	24	2	1
Работа	1	1	2	1	1	2	1	2	2	4	17	1	2
Хорошо	2	4	4	2	3	1	2	4	3	2	27	3	3
Лицо	3	8	8	5	5	5	4	6	8	6	58	5	4
Друг	5	5	3	4	4	6	8	3	1	8	47	4	5
Длинный	6	3	5	7	7	8	7	5	7	5	60	6	6
Характер	7	6	6	10	6	4	5	8	5	3	60	7	7
Сигнал	10	10	9	6	10	10	10	10	6	10	91	10	8
Неизвестный	8	9	10	8	8	9	9	9	9	7	86	9	9
Энергичный	9	7	7	9	9	7	6	7	10	9	80	8	10

Мы не будем сейчас рассматривать вопрос о степени близости объективного и субъективного ранжирования, а обратим внимание читателя на то, что два слова — *длинный* и *характер* — получили одинаковое количество баллов. Эта ситуация встречается довольно часто при построении вариационных рядов. Примером могут служить частотные словари, в которых большие массивы редких словоформ имеют одинаковые частоты [33, с. 376—567]. В этом случае упорядочение происходит либо по какому-либо качественному признаку, например по алфавитному, либо обеим вариантам приписывается одинаковый ранг, представляющий собой среднее арифметическое порядковых номеров, либо ранг первой варианты в группе наших вариант.

4. Качественный признак у лингвистических единиц. Качественными признаками группировки вариант являются такие признаки, которые не содержат ни количественной оценки вариант, ни возможности их ранжирования. Примером может служить группировка словоформ по семантическим или грамматическим классам, или расположение фонем, исходя из иерархии дифференциальных

признаков [24, с.46]. В этих случаях группировка вариант, отобранных по качественному признаку, заключается в их классификации по градации этого признака.

Например, при статистическом исследовании румынских текстов по радиоэлектронике выделены группы именных словоформ, причем упорядочение этих групп осуществлено по убыванию их номинативности (см. табл. 7.13).

Таблица 7.13

Статистика именных частей речи в румынских текстах по радиоэлектронике

	Количество разных словоформ в частотном словаре	Количество словоупотреблений в тексте	
		F	$\frac{F}{N} \cdot 100\%$
Существительные (С)	7180	68781	59,1
Местоимения (М)	117	11124	9,4
Числительные (Ч)	111	2602	2,5
Прилагательные (П)	2216	19654	16,3
Адъектированные причастия (Ад)	1262	4696	4,1
Артикли (Ар)	20	10572	8,6
Всего	10906	117429	100,0

Можно группировать лингвистические элементы по альтернативному качественному признаку, т. е. по наличию (A) или по отсутствию ($\overline{A}$) какого-либо признака. Например, в только что рассмотренном примере таким признаком может быть принадлежность словоформы к именным частям речи. Таких словоформ в румынском частотном словаре по электронике 10 906, и они дадут 117 429 словоупотреблений в общей выборке из 200 тыс. словоупотреблений. Соответственно 3 386 словоформ словаря и 82 571 словоупотребление текста не будут обладать этим признаком.

5. Графическое изображение лингвистических вариационных рядов. Слабой стороной табличного описания колебания признака является недостаточная наглядность этого описания. Гораздо большей наглядности мы можем достичь с помощью графического или геометрического изображения интересующего нас распределения. Ведь геометрическая интерпретация математической зависимости не только придает им наглядность, но позволяет также анализировать эти зависимости в простой и доступной форме. В лингвистике чаще всего применяются такие формы графического представления, как полигон, гистограмма, кумулятивная кривая и огива, а также диаграмма. Все указанные графики строятся в прямоугольной системе координат. Масштабы на осях ординат и абсцисс могут быть произвольными: обычно их выбирают так, чтобы соотношение ширины и высоты графика было равно 1:2.

Теперь рассмотрим каждое из перечисленных графических представлений.

1. *Многоугольником распределения признака* (*полигоном*) называют замкнутую ломаную линию, соединяющую вершины ординат, соответствующих частотам (или частостям) каждого из значений признака. При построении полигона в начале и в конце вариационного ряда добавляются нулевые варианты (варианты, частоты которых равны нулю). Это дает возможность, соединив указанные точки, замкнуть ломаную линию, которая и образует полигон. На рис. 48 полигон распределения построен для дискретного, не сгруппированного в интервалы значения признака. Можно построить полигон и для сгруппированного интервала. Для этого берется середина интервала и от нее строится ордината, соответствующая сумме вариант, относящихся к данному интервалу (рис. 49). Если же распределение непрерывно, то от середины интервала берется ордината, соответствующая частоте (или частости; рис. 50). Как и в случае с дискретным признаком, в начале и в конце ряда добавлены нулевые варианты.

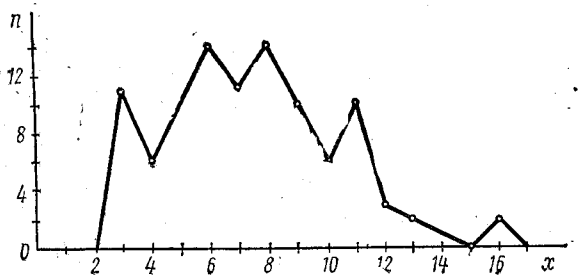


Рис. 48

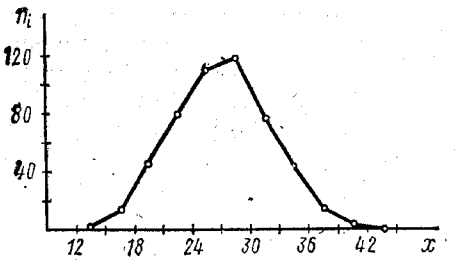


Рис. 49

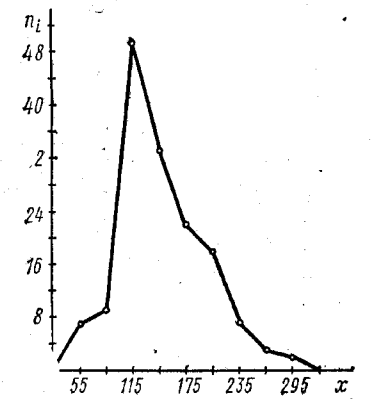


Рис. 50

2. При графическом изображении интервальных вариационных рядов обычно используется не полигон, а так называемая *гистограмма*. Отличие гистограммы от полигона состоит в том, что на оси аб-

сцисс откладываются не точки, а отрезки, соответствующие ширине интервала. Полученные отрезки являются основаниями прямоугольников, высоты которых пропорциональны частотам (или относительным частостям) соответствующих интервалов. В итоге получается ступенчатая фигура, образованная рядом сдвинутых прямоугольников (рис. 51).

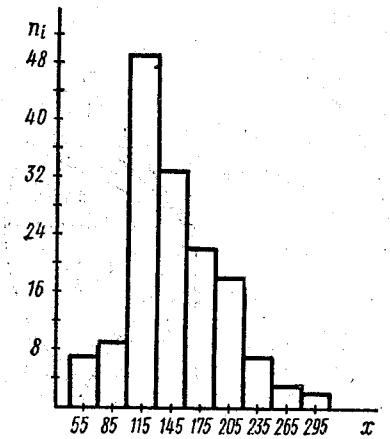


Рис. 51

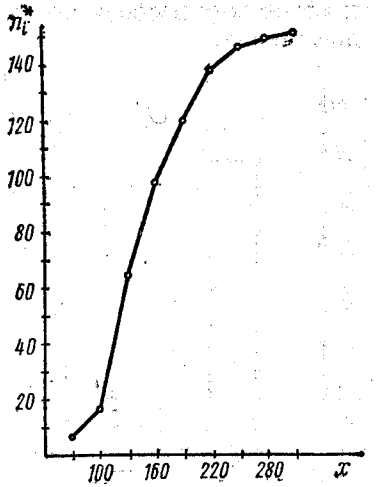


Рис. 52

Полигон и гистограмма представляют собой два крайних случая идеализации интервальных вариационных рядов: если при построении полигона частот все значения, лежащие внутри интервала, «стягиваются» к его середине, то при использовании гистограммы они представляются равномерно распределенными по всему интервалу. Полигон используется тогда, когда статистическая совокупность существенно дискретна. Гистограмма используется чаще для изображения непрерывных вариационных рядов.

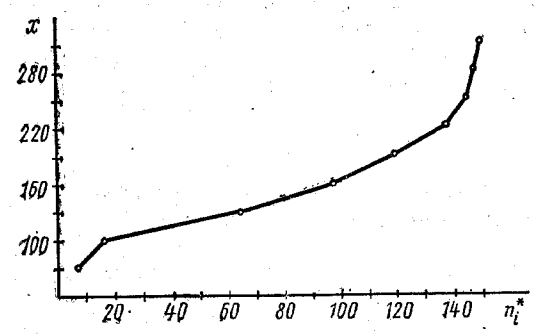


Рис. 53

3. *Кумулятивной кривой* называется ломаная линия, которая соединяет вершины ординат, соответствующие накопленным частотам для данного значения признака (*накопленной частотой* называется сумма частоты данной варианты и частот всех предыдущих вариант). На рис. 52 показана кумулятивная кривая накопленных частот длин китайских слогов, показанных в табл. 7.9.

4. Если поменять ролями оси координат и откладывать на оси абсцисс не значения признака, а накопленные частоты (или относительные частоты), а на оси ординат — варианты (значения признака) распределения, то получим так называемую *огиву* распределения (рис. 53). И огива, и кумулятивная кривая используются для графического изображения как дискретных, так и непрерывных распределений.

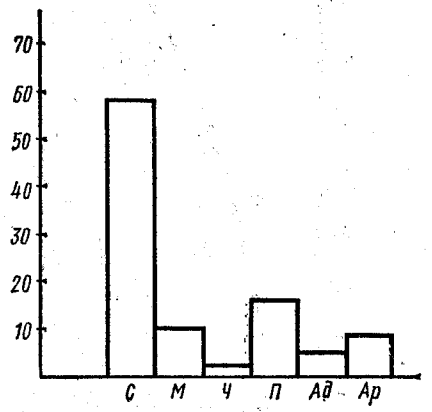


Рис. 54

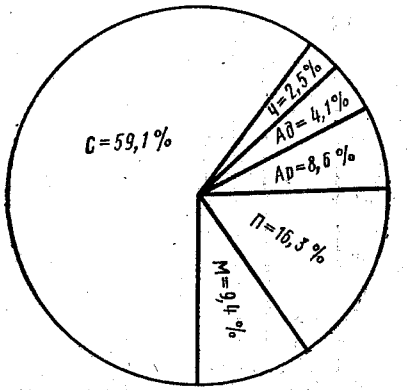


Рис. 55

5. Порядковый признак не требует графической интерпретации. Что же касается качественных признаков, то для их графической интерпретации используются столбиковые и круговые *диаграммы*. Такие диаграммы, построенные на основании данных табл. 7.13 и характеризующие удельные веса румынских именных форм, приведены на рис. 54 и 55. Графической интерпретацией вариаций качественного лингвистического признака можно считать и карту лингвистического атласа.

§ 3. Статистические характеристики лингвистических вариационных рядов

Для того чтобы иметь возможность сравнивать вариационные ряды с вероятностными моделями, рассмотренными в гл. 6, необходимо выявить набор тех характеристик (параметров), которые представляли бы в обобщенном виде основные свойства данной лингвостатистической совокупности. Наиболее важными статистическими характеристиками эмпирических распределений считаются с р е д н и е (средняя арифметическая, средняя геометрическая, средняя гармоническая, медиана, мода, квартили, децили и т. д.), а также м е р ы р а с с е и в а н и я (размах, абсолютное отклонение, среднее квадратическое отклонение, опытная дисперсия, коэффициент вариации, квартильное отклонение и др.).

1. Средняя арифметическая. Среди разных видов средних наиболее широкое применение в лингвистических работах имеет *средняя арифметическая*. Она проще других и по смыслу, и по свойствам, и по способу получения. Средняя арифметическая ($\bar{x}$) признака есть отношение суммы количественных вариант этого признака к общему числу вариант, т. е.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum_{j=1}^{N} x_j}{N} = \frac{1}{N} \sum_{j=1}^{N} x_j. \qquad (7.3)$$

Например, для исследования распределения букв, передающих гласные, из русского газетного текста («Правда» от 29.X.69 г.) извлечено 20 порций по 10 букв в каждой. При этом получен следующий неупорядоченный ряд появления гласных в каждой порции:

5, 4, 3, 3, 4, 4, 3, 4, 4, 4,
4, 6, 5, 4, 3, 5, 3, 3, 3, 4.

Средняя арифметическая частоты появления гласной в газетном сегменте длиной в 10 букв, вычисленная с помощью формулы (7.3), составляет

$$\bar{x}_{\text{гл}} = \frac{5+4+3+3+4+4+3+4+4+4+4+6+5+4+3+5+3+3+3+4}{20} =$$

$$= \frac{78}{20} = 3{,}9.$$

В тех случаях, когда мы имеем дело со сгруппированным вариационным рядом (см. табл. 7.2), среднюю арифметическую можно определить как отношение сумм произведений вариант на их веса к сумме весов:

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{N} = \frac{1}{N} \sum_{i=1}^{k} n_i x_i \qquad (7.4)$$

Если учесть, что отношения $\frac{n_i}{N} = f_i$ являются частостями (относительными частотами) вариант, то средняя арифметическая может быть также получена как сумма произведений вариант на их частости:

$$\bar{x} = \sum_{i=1}^{k} f_i x_i. \qquad (7.5)$$

Среднюю арифметическую, выраженную формулами (7.4) и (7.5), в дальнейшем будем называть *средневзвешенной*.

Представим данные о распределении гласных в русском газетном тексте (см. выше) в виде вариационного ряда

x_i	3	4	5	6	N
n_i	7	9	3	1	20

(где $k = 4$); следовательно,

$$x_{\text{гл}} = \frac{7 \cdot 3 + 9 \cdot 4 + 3 \cdot 5 + 1 \cdot 6}{20} = \frac{21 + 36 + 15 + 6}{20} = \frac{78}{20} = 3{,}9.$$

2. Свойства средней арифметической. Средняя арифметическая характеризуется свойствами, некоторые из которых совпадают со свойствами математического ожидания (см. гл. 6, § 2, п. 8).

1. *Средняя арифметическая постоянной равна этой постоянной*, т. е.

$$\bar{x} = a \text{ при } x_1 = x_2 = \ldots = x_i = \ldots = x_N = a.$$

2. *Сумма отклонений вариант от среднего значения равна нулю*, т. е.

$$\sum_{i=1}^{k} n_i (x_i - \bar{x}) = 0.$$

3. *Если увеличить (уменьшить) значение признака на постоянное число a, то средняя арифметическая также увеличится (уменьшится) на это число*, т.е.

$$\frac{1}{N} \sum_{i=1}^{k} n_i (x_i + a) = \bar{x} + a.$$

4. *Если увеличить (уменьшить) каждое значение признака в a раз, то средняя арифметическая также увеличится (уменьшится) в a раз*, т. е.

$$\frac{1}{N} \sum_{i=1}^{k} n_i (a x_i) = a\bar{x}.$$

5. *Средняя арифметическая суммы признаков равна сумме средних арифметических*, т. е. если признак X является суммой признаков Y и Z, то

$$\bar{x} = \bar{y} + \bar{z}.$$

Среднее количество гласных звукотипов в одной десятибуквенной порции газетного текста (см. выше) можно получить, зная среднюю арифметическую для букв, которые передают гласные, стоящие после твердых согласных [ы], [э], [а], [о], [у], и среднюю арифметическую для звукотипов, находящихся после мягких согласных [и], [е], [ä], [ë], [ÿ]. Неупорядоченный ряд для гласных первого типа (будем считать их признаком Y) имеет вид

3	2	1	2	3	2	1	2	3	2
2	3	3	2	2	3	2	2	2	3

Неупорядоченный ряд для гласных второго типа (признак Z) выглядит так:

2	2	2	1	1	2	2	2	1	2
2	3	2	2	1	2	1	1	1	1

Применяя формулу (7.3), имеем

$$\bar{y} = 45/20 = 2{,}25,\ \bar{z} = 33/20 = 1{,}65,$$

а на основании свойства средней арифметической можно записать

$$\bar{x} = \bar{y} + \bar{z} = 2{,}25 + 1{,}65 = 3{,}9,$$

что соответствует результату, полученному ранее.

С помощью метода индукции это свойство распространяется и на те случаи, когда признак X представляет сумму (или разность) трех и более признаков.

3. Вычисление средней арифметической с помощью метода моментов. Упростить ход вычисления средней арифметической можно с помощью метода моментов. Первая формула этого метода имеет вид

$$\bar{x} = a + \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - a). \qquad (7.6)$$

Здесь a — произвольное постоянное число, подбираемое обычно так, чтобы разности $x_i - a$ были бы возможно проще и меньше (это достигается тогда, когда величина a занимает серединное положение в данном ряде и имеет наибольший вес).

Пользуясь соотношением (7.6), произведем вычисление средней арифметической в вариационном ряде распределения гласных в русском газетном тексте (см. выше). Взяв $a = 4$, составим табл. 7.14.

Таблица 7.14

x_i	n_i	$x_i - a$	$\frac{1}{n_i (x_i - a)}$
3	7	−1	−7
4	9	0	0
5	3	1	3
6	1	2	2
Сумма	20		−2

Отсюда $\bar{x} = 4 - 2/20 = 3{,}9$.

В тех случаях, когда либо все частоты n_i, либо все разности $x_i - a$ имеют общий множитель l, для упрощенного вычисления

средней арифметической используется формула моментов (ср. § 4), построенная на использовании свойств 3 и 4 средней арифметической:

$$\bar{x} = a + \frac{l}{N}\sum_{i=1}^{k} \frac{n_i (x_i - a)}{l}. \qquad (7.7)$$

Воспользовавшись данными интервального ряда, приведенного в табл. 7.9, вычислим среднюю арифметическую с помощью формулы моментов. В качестве x_i возьмем середины интервальных рядов (обозначим их через $\hat{x}_i$) и положим $a = 145$. Поскольку разности $\hat{x}_i - a$ всегда кратны тридцати, примем $l = 30$. Все эти данные приведены в табл. 7.15.

Таблица 7.15

Интервалы длин слов (в мс)	Середина интервала $\hat{x}_i$	n_i	$\hat{x}_i - a$	$\frac{\hat{x}_i - a}{l} n_i$
40—70	55	7	−90	−21
70—100	85	9	−60	−18
100—130	115	49	−30	−49
130—160	145	33	0	0
160—190	175	22	30	22
190—220	205	18	60	36
220—250	235	7	90	21
250—280	265	3	120	12
280—310	295	2	150	10
Сумма		150		13

Используя формулу (7.7), видим, что средняя арифметическая длин китайских слогов равна

$$\bar{x} = 145 + \frac{30 \cdot 13}{150} = 147{,}6 \text{ (мс)}.$$

4. Степенные средние. В статистике используются различные виды средних, обобщаемые формулой

$$\bar{x}_\alpha = \left(\sum_{i=1}^{k} \frac{n_i x_i^\alpha}{N}\right)^{1/\alpha}. \qquad (7.8)$$

Если $\alpha = 1$, то выражение (7.8) превращается в формулу средней арифметической (7.4).

При $\alpha = 2$ выражение (7.8) становится формулой *средней квадратической*

$$\bar{x}_2 = \sqrt{\sum_{i=1}^{k} \frac{n_i x_i^2}{N}}.$$

Аналогичным путем можно получить $\bar{x}_3$, $\bar{x}_4$ и т. д.

Если взять $\alpha = -1$, то получаем *среднюю гармоническую*, вычисляемую по формуле

$$\bar{x}_m = \frac{N}{n_1/x_1 + n_2/x_2 + \ldots + n_k/x_k} = \frac{N}{\sum_{i=1}^{k} n_i/x_i}.$$

5. Медиана, квартили и децили. Значение признака, относительно которого совокупность делится пополам, называется *медианой* (Me). Если имеется нечетное число значений варьирующего признака $x_1, x_2, \ldots, x_{m-1}, x_m, x_{m+1}, \ldots, x_{2m-1}$, которые расположены в возрастающем порядке, то медианой этого распределения служит варианта x_m. Слева и справа от значения $\text{Me} = x_m$ расположено по $m - 1$ значений признака.

Например, для анализа лингвистических терминологических систем [7, с. 49] взято семь выборок из русских лингвистических текстов (объем каждой выборки 250 терминоупотреблений). После подсчета в каждой выборке числа употреблений слова *лицо* получена следующая упорядоченная совокупность числовых вариант употреблений указанного слова:

x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	1	3	4	9	10	12

Определим медиану ряда. Здесь нечетное число вариант, равное 7; поэтому $7 = 2m - 1$, откуда $m = 4$. Таким образом, $\text{Me} = x_4 = 4$.

В тех случаях, когда имеется четное число вариант $x_1, x_2, \ldots, \ldots, x_{m-1}, x_m, x_{m+1}, \ldots, x_{2m}$, медиана, делящая совокупность на две равные части по m членов, находится между признаками x_m и x_{m+1}. Значение медианы условно определяется как средняя арифметическая двух вариант, находящихся в середине ряда:

$$\text{Me} = (x_m + x_{m+1})/2.$$

Например, при исследовании тех же лингвистических терминологических систем в русском языке получена следующая упорядоченная совокупность употребления термина *значение* в десяти выборках:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	7	8	8	8	10	10	13	16	20

Найдем медиану ряда. Здесь число вариант четное: $2m = 10$, а $m = 5$, откуда находим $\text{Me} = (8 + 10)/2 = 9$.

Несколько сложнее вычисление медианы в случае интервального вариационного ряда. Здесь сначала выявляют *медианный интервал*,

т. е. тот интервал, в котором находится медиана. Затем вычисляют величину медианы, пользуясь формулой

$$\text{Me} = X_{\text{Me (min)}} + k \frac{\frac{N}{2} - S_{\text{Me}-1}}{n_{\text{Me}}}, \qquad (7.9)$$

где $X_{\text{Me(min)}}$ — нижняя граница медианного интервала; $S_{\text{Me}-1}$ — накопленная частота, которая соответствует интервалу, предшествующему медианному; n_{Me} — частота медианного интервала; k — ширина медианного интервала (интервальная разность); N — объем всей совокупности.

Пользуясь данными интервального вариационного ряда, приведенного в табл. 7.9, вычислим медиану распределения длин китайских слогов. Для этого перепишем сначала эту таблицу, добавив данные о серединах интервалов, а также о накопленных частотах и частостях (табл. 7.16).

Таблица 7.16

Номер интервала	Интервалы (в мс)	Частота	Накопленная частота	Накопленная частость
1	40—70	7	7	0,047
2	70—100	9	16	0,107
3	100—130	49	65	0,433
4	130—160	33	98	0,653
5	160—190	22	120	0,800
6	190—220	18	138	0,920
7	220—250	7	145	0,967
8	250—280	3	148	0,987
9	280—310	2	150	1,000

Нетрудно заметить, что четвертый интервал является медианным, отсюда имеем: $X_{\text{Me(min)}} = 130$, $S_{\text{Me}-1} = 65$, $n_{\text{Me}} = 33$, $k = 30$, $N = 150$.

Подставив эти значения в формулу (7.9), получаем

$$\text{Me} = 130 + 30 \cdot \frac{75 - 65}{33} = 139{,}09.$$

Медиана делит ранжированный ряд на две равные группы. Аналогичным образом можно найти такие значения признака, которые делят каждую группу снова на две равные части. Эти значения признака называются *квартилями*. Различают три квартили: *первую* (или *нижнюю*) *квартиль* (Q_1), делящую пополам ту часть ранжированного ряда, которая находится ниже медианы; *вторую квартиль*, совпадающую с медианой, *третью* (или *верхнюю*) *квартиль* (Q_3), которая делит пополам ту часть ряда, которая расположена выше медианы.

Те значения признака, которые делят объем исследуемой совокупности на десять равных по объему групп, называются *децилями*.

6. Мода. *Модой* (Мо) называется наиболее часто встречающаяся варианта данного вариационного ряда (ср. гл. 6, § 1, п. 3). Определение моды для дискретного распределения не представляет трудности. Модой здесь служит та варианта, которой соответствует наибольшая частота (кстати, таких вариант может быть несколько).

Несколько сложнее определить моду в интервальных непрерывных распределениях. Здесь, как и при вычислении медианы, начинают с определения *модального интервала*, т. е. того интервала, внутри которого находится мода (таких интервалов может быть несколько). Затем численное значение моды определяют по следующей приближенной формуле:

$$\text{Мо} = X_{\text{Мо (min)}} + k \frac{n_{\text{Мо}} - n_{\text{Мо}-1}}{(n_{\text{Мо}} - n_{\text{Мо}-1}) + (n_{\text{Мо}} - n_{\text{Мо}+1})}, \quad (7.10)$$

где $X_{\text{Мо(min)}}$ — нижняя граница модального интервала; k — длина модального интервала; $n_{\text{Мо}}$ — частота модального интервала; $n_{\text{Мо}-1}$ — частота интервала, предшествующего модальному; $n_{\text{Мо}+1}$ — частота интервала, следующего за модальным.

Пользуясь данными табл. 7.9, вычислим моду распределения длин китайских слогов. Модальным здесь является третий интервал; следовательно,

$$\text{Мо} = 100 + 30 \frac{49-9}{(49-9)+(49-33)} = 121{,}43.$$

7. Соотношение между средней арифметической, медианой и модой. Соотношение между этими тремя основными параметрами эмпирических распределений используется для оценки асимметрии распределений. Нетрудно заметить, что в тех случаях, когда распределение симметрично, выполняется равенство $\overline{x} = \text{Ме} = \text{Мо}$.

В случае умеренной асимметрии вариационного ряда — явлении, часто встречающемся в лингвистике, — имеет место следующее приближенное равенство: $\text{Мо} \approx \overline{x} - 3(\overline{x} - \text{Ме})$, иными словами, *медиана расположена между модой и средней арифметической так, что расстояние от нее до моды равно двум расстояниям от медианы до средней арифметической*. В случае умеренно скошенных распределений этим соотношением пользуются для грубой оценки неизвестного параметра (скажем, моды) по двум известным характеристикам, например медианы и средней арифметической [7, с.55].

Сделаем теперь несколько замечаний об использовании указанных статистических параметров. Из всех этих параметров наиболее простой по смыслу и по способу получения является средняя арифметическая. В отличие от моды и медианы средняя арифметическая легко поддается аналитическим операциям: выше уже указывалось, что при объединении двух распределений с различными средними средняя полученного распределения равна сумме средних из отдельных распределений. Поэтому, если нет существенных доводов в пользу иного вида средней, следует пользоваться средней арифметической.

Вместе с тем следует помнить, что средняя арифметическая изменяется с изменением значения любого признака. Особенно она чувствительна к колебаниям крайних вариант распределений. Иначе обстоит дело с медианой: из ее определения следует, что медиана не зависит от значений признаков, лежащих справа и слева от нее (важно лишь, чтобы число признаков, меньших и больших, чем медиана, оставалось неизменным). Поэтому медиану целесообразно использовать в качестве средней для таких распределений, концы которых определены недостаточно надежно.

Что касается моды, то она служит средством выявления одного или нескольких значений признака, около которых группируется большая часть объема асимметричного лингвистического распределения. В лингво-статистике модальные характеристики распределения могут быть использованы для объективного выделения терминологических, ключевых и вообще доминантных слов и словосочетаний текста [32а, с.47—112].

8. Рассеяние значений признака. Размах вариации. Хотя средняя арифметическая, мода, медиана и другие средние признаки дают ориентировочную количественную характеристику лингвистической единицы, они не учитывают степень равномерности употребления этой единицы в текстах. Между тем учет количественной вариации лингвистического признака в изучаемом тексте имеет принципиальное значение для языковеда. Всякая вариация лингвистической случайной величины передает в конечном итоге лексические, грамматические, стилевые и другие внутрилингвистические и экстралингвитические особенности текста.

Наиболее простой характеристикой рассеивания признака является *размах вариации* R, который определяется разностью

$$R = x_{\max} - x_{\min}.$$

Рассмотрим в этой связи два вариационных ряда частот немецкого существительного Kraft в двух выборках публицистических текстов — одна из газет ГДР (табл. 7.17), другая — из газет ФРГ (табл. 7.18). Каждая выборка состоит из 20 текстов по 1000 словоупотреблений каждый [7, с. 57].

Таблица 7.17

x_i	0	1	2	3	4	5	6	7
n_i	16	2	0	0	1	1	0	0

Таблица 7.18

x'_i	0	1	2	3	4	5	6	7
n'_i	9	11	0	0	0	0	0	0

Используя формулу (7.4), находим, что средние арифметические значения для обоих рядов вариант одинаковы:

$$\bar{x} = \bar{x}' = 0{,}55.$$

Вместе с тем эти ряды дают различный размах вариации:

$$R(x) = 16 - 0 = 16,\ R(x') = 11 - 0 = 11.$$

Однако размах вариации является очень приближенной оценкой степени рассеивания признака, так как совершенно не учитывает положений и «весов» вариант признака, находящихся в пределах крайних вариант. Действительно, хотя размах вариации немецкого существительного Kraft в текстах ГДР выше, чем в западногерманских текстах, крайние варианты в первом случае встречаются редко и имеют малый вес, поэтому вряд ли можно уверенно говорить о том, что рассеяние здесь действительно выше, чем в текстах ФРГ.

9 Линейное отклонение. Более точную оценку рассеивания можно получить, учитывая абсолютные величины отклонений $|x_i - \bar{x}|$ значения признака от его средней арифметической. Среднее значение этих абсолютных величин, называемое *линейным отклонением*, вычисляется для несгруппированного вариационного ряда по формуле

$$\overline{|x_i - \bar{x}|} = \frac{1}{N}\sum_{i=1}^{N} |x_i - \bar{x}|,$$

а для сгруппированного ряда — по формулам

$$\overline{|x_i - \bar{x}|} = \frac{1}{N}\sum_{i=1}^{k} n_i |x_i - \bar{x}| \qquad (7.11)$$

или

$$\overline{|x_i - \bar{x}|} = \sum_{i=1}^{k} f_i |x_i - \bar{x}|. \qquad (7.12)$$

Линейные отклонения для вариационных рядов существительного Kraft (см. табл. 7.17 и 7.18) соответственно составляют

$$\overline{|x_i - \bar{x}|} = 0{,}88 \text{ и } \overline{|x_i' - \bar{x}|} = 0{,}50.$$

Статистическая вариация в употреблении существительного Kraft в публицистических текстах ГДР несколько выше, чем в газетных текстах ФРГ, однако различия в рассеивании здесь не столь значительны, как при оценке по размаху вариации.

Нетрудно заметить, что линейное отклонение имеет ту же размерность, что и величина средней арифметической данного вариационного ряда. Поэтому если два вариационных ряда имеют разные значения $\bar{x}$, то их линейные отклонения оказываются несопоставимыми величинами. В том случае, когда возникает необходимость в численном сравнении вариаций в распределениях разных лингвистических признаков, необходимо привести эти вариации к некоторому «общему знаменателю». Это достигается путем применения так называемого *коэффициента вариации*, представляющего собой средний процент рассеивания значений случайной величины по отношению к средней арифметической:

$$V = \frac{\overline{|x_i - \bar{x}|}}{\bar{x}} \cdot 100\ \%. \qquad (7.13)$$

С помощью выражения (7.13) можно показать, что для существительного Kraft значения коэффициента вариации соответственно равны $V_x = 160\%$ и $V_{x'} = 90\%$.

Рассмотрим еще один вариационный ряд. Этот ряд (см. табл. 7.19) отражает распределение частот английского определенного артикля the в десяти английских научных текстах по 1000 словоупотреблений каждый [7, с.62].

Таблица 7.19

Номера текстов	11	46	13	2	3; 25	47	43	9	1
x_i	67	68	71	72	74	80	82	83	84
f_i	0,1	0,1	0,1	0,1	0,2	0,1	0,1	0,1	0,1

С помощью формул (7.5) и (7.12) получаем, что средняя арифметическая этого ряда составляет $\bar{x} = 75{,}5$, а линейное отклонение равно $\overline{|x_i - \bar{x}|} = 5{,}4$; следовательно, коэффициент вариации

$$V = \frac{5{,}4}{75{,}5} \cdot 100\ \% = 7{,}12\%.$$

Легко заметить, что коэффициент вариации у английского артикля заметно меньше, чем коэффициент вариации у немецкого существительного. Это неудивительно: служебные формы обычно имеют во всех языках менее рассеянное употребление, чем знаменательные слова.

10. Опытная дисперсия и стандарт. Линейное отклонение не всегда улавливает истинную закономерность вариации случайной величины, так как результаты здесь сильно усредняются и сглаживаются, а большие отклонения становятся мало ощутимыми, особенно при большом числе испытаний. Между тем при решении ряда лингвистических и особенно инженерно-лингвистических задач учет именно больших отклонений оказывается принципиально важным. Чтобы учесть долю больших отклонений, рассматривают не сами отклонения, а их квадраты.

Сумма взвешенных квадратов отклонения вариант от среднего арифметического, называемая *опытной дисперсией* (или просто *дисперсией*), для несгруппированного ряда подсчитывается по формуле

$$D = s^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2.$$

Для сгруппированного ряда дисперсия определяется по формуле

$$D = s^2 = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2 = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2. \qquad (7.14)$$

Размерность дисперсии равна квадрату размерности вариант. Чтобы вернуться к мере рассеивания, имеющей тот же порядок, что и сами варианты, а также их отклонения, вводят новую характеристику — *стандарт*, или *выборочное среднее квадратическое отклонение*, равное квадратному корню из дисперсии:

$$s = \sqrt{D} = \sqrt{\sum_i f_i (x_i - \bar{x})^2}. \qquad (7.15)$$

Если же нужно сопоставить рассеяние разных по качеству признаков, оцененное с помощью стандарта, используется *коэффициент вариации*

$$V(s) = \frac{s}{\bar{x}} \cdot 100\%. \qquad (7.16)$$

Для иллюстрации определим дисперсию, выборочное квадратическое отклонение и коэффициент вариации в распределениях частот английского артикля the и немецкого существительного Kraft (данные приведены в табл. 7.17—7.19). Значения коэффициента вариации по стандарту $V(s)$ сравним со значениями коэффициента вариации V, полученными по абсолютному отклонению.

Найденные с помощью формул (7.14) и (7.16) величины приведены в табл. 7.20.

Таблица 7.20

Словоформы	$D_x = s^2$	$s = \sqrt{D_x}$	$V(s)$ (в %)	V (в %)
the	39,70	6,30	8,34	7,5
Kraft (тексты ГДР)	1,85	1,36	247,4	160,0
Kraft (тексты ФРГ)	0,25	0,50	90,0	90,0

Рассеяние контрольных словоформ, оцененное с помощью среднего квадратического, в целом соответствует рассеянию, полученному по линейному отклонению. Однако поскольку стандарт учитывает то влияние, которое оказывает на конечный результат рассеяние крайних вариант, значение коэффициента вариации $V(s)$ больше значения V.

11. Свойства опытной дисперсии. Основные свойства опытной дисперсии совпадают со свойствами теоретической дисперсии.

1. *Дисперсия постоянной величины равна нулю:*

$$D(C) = 0. \qquad (7.17)$$

2. *Постоянную можно вынести за знак дисперсии, возведя ее в квадрат:*

$$D(CX) = C^2 D_x. \qquad (7.18)$$

3. *Увеличение (уменьшение) значений признака на одну и ту же постоянную C не изменяет дисперсии:*

$$D(X \pm C) = D_x. \qquad (7.19)$$

4. *Дисперсия равна средней арифметической квадратов значений признака без квадрата их средней арифметической:*

$$D_x = \frac{\sum n_i x_i^2}{N} - (\bar{x})^2. \qquad (7.20)$$

Проиллюстрируем это свойство на примере распределения частот английского артикля the. Для этого воспользуемся столбцами (1)—(4) табл. 7.21.

Таблица 7.21

x_i	n_i	$n_i x_i$	$n_i x_i^2$	$x_i - a$	$n_i (x_i - a)$	$n_i (x_i - a)^2$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
67	1	67	4489	—7	—7	49
68	1	68	4624	—6	—6	36
71	1	71	5041	—3	—3	9
72	1	72	5184	—2	—2	4
74	2	148	10952	0	0	0
80	1	80	6400	6	6	36
82	1	82	6724	8	8	64
83	1	83	6889	9	9	81
84	1	84	7056	10	10	100
Суммы	10	755	57359			379

Подставляя величины из нижней строки табл. 7.21 в формулы (7.4) и (7.14), имеем

$$\bar{x} = \frac{755}{10} = 75{,}5; \quad (\bar{x})^2 = 5700{,}25; \quad \frac{\sum n_i x_i^2}{N} = \frac{57359}{10} = 5735{,}9;$$

$$D_x = 5735{,}9 - 5700{,}25 = 35{,}65, \quad s = \sqrt{35{,}65} = 5{,}97.$$

5. *Дисперсия признака относительно средней арифметической равна дисперсии признака относительно произвольной величины (a) минус квадрат разности между средней арифметической и этой величиной:*

$$D_x = \frac{1}{N}\sum_i n_i (x_i - a)^2 - (\bar{x} - a)^2. \qquad (7.21)$$

Обычно нахождение дисперсии и среднего квадратического отклонения непосредственно с помощью выражений (7.14) и (7.15) связано с трудоемкими вычислениями. Использование свойства 5 дисперсии значительно упрощает процедуру вычисления, которая оказывается аналогичной нахождению средней арифметической по методу моментов.

Проиллюстрируем этот прием вычисления дисперсии снова на примере распределения частот артикля the. Используя столбцы (5)—(7) табл. 7.21 и полагая $a = 74$, согласно формуле (7.21) получаем

$$D = \frac{1}{10} \cdot 379 - (75{,}5 - 74)^2 = 35{,}65,$$

откуда $s = \sqrt{35{,}65} = 5{,}97$.

12. Средняя арифметическая и дисперсия для нескольких совокупностей. До сих пор мы имели дело со средней арифметической и дисперсией, характеризовавшими одну совокупность. Однако на практике постоянно встречаются случаи, когда та или иная лингвистическая совокупность образуется в результате соединения нескольких частных совокупностей с одним и тем же признаком, но с разными его распределениями и, следовательно, с различными средними арифметическими и дисперсиями.

Каждую из этих самостоятельных совокупностей мы будем называть *частной совокупностью*. Характеризующую каждую частную совокупность среднюю арифметическую признака назовем *внутренней* (или *групповой*) *средней* ($\bar{x}_i$), а соответствующие частные дисперсии определим как *внутренние* (или *групповые*) *дисперсии* ($D_{r_i} = s_{r_i}^2$).

Полученная в результате объединения нескольких частных совокупностей общая совокупность имеет свою общую среднюю арифметическую или просто общую среднюю $\bar{x}$. Вычисление общей средней производится согласно теореме сложения средних: *если статистическая совокупность S состоит из S_1, S_2, ..., S_m частных совокупностей объемом l каждая, то общая средняя равна средней арифметической внутренних средних*, т. е.

$$\bar{x} = \frac{\bar{x}_1 + \bar{x}_2 + \ldots + \bar{x}_m}{m}, \qquad (7.22)$$

где $\bar{x}_1$, $\bar{x}_2$, ..., $\bar{x}_m$ — средние арифметические частных совокупностей.

Только что приведенная теорема описывает тот частный случай, когда объемы частных совокупностей одинаковы и равны l. Эта теорема легко доказывается и в том случае, когда объемы частных совокупностей различны. Если эти объемы соответственно составляют

$l_1, l_2, \ldots, l_m$, то общая средняя равна средней из соответственно взвешенных частных средних. Иными словами,

$$x = \frac{l_1 \bar{x}_1 + l_2 \bar{x}_2 + \ldots + l_m \bar{x}_m}{l_1 + l_2 + \ldots + l_m}, \qquad (7.23)$$

или короче

$$\bar{x} = \frac{\sum_{i=1}^{m} l_i \bar{x}_i}{\sum_{i=1}^{m} l_i}.$$

Например, в результате статистического описания классических и позднелатинских текстов [27, с. 53—54] получены данные о соотношении препозитивного и постпозитивного употребления указательного местоимения ille при определяемом существительном в различных жанрах (табл. 7.22).

Таблица 7.22

Статистика препозитивного употребления ille в текстах классической и поздней латыни

Жанр	Авторы и памятники	Количество препозитивных ille	Общее число учтенных присубстантивных ille	Внутренняя средняя x(%) препозитивных ille	Вес жанра (число тысяч употреблений) ille
Эпистолярные тексты	1. Цицерон «Письма»	47	50		
	2. Плиний «Письма»	27	97		
	3. Кассиодор «Сочинения»	18	21		
	Всего по жанру	92	168	54,8	0,168
Повествовательная проза	1. Цезарь «Записки о галльской войне»	23	25		
	2. Тацит «Анналы»	9	10		
	3. «Путешествие Этерии»	74	108		
	4. «История франков»	22	49		
	5. «Салическая Правда»	32	56		
	Всего по жанру	160	248	64,5	0,248
Ораторский стиль	1. Цицерон «Речи»	1735	2155		
	2. Сенека «О благодеянии»	6	8		
	3. Св. Августин «Исповедь»	110	159		
	Всего по жанру	1851	2322	79,7	2,322

Подставляя данные из таблицы в формулу (7.23), получаем

$$\bar{x} = \frac{54{,}8 \cdot 168 + 64{,}5 \cdot 248 + 79{,}7 \cdot 2322}{168 + 248 + 2322} = 76{,}8\%,$$

т. е. около 77% препозитивных ille.

При исследовании рассеяния в нескольких лингвистических совокупностях используются следующие понятия:

1) *внутренняя дисперсия*

$$D_{r_i} = s_{r_i}^2 = \frac{1}{l} \sum_{k=1}^{l} (x_{ik} - \bar{x}_i)^2, \qquad (7.24)$$

где i — номер частной совокупности, l — число вариант признака в этой совокупности, а k — номер признака [ср. с формулой (7.14)];

2) *общая дисперсия*

$$D = s^2 = \frac{1}{ml} \sum_{k=1}^{l} \sum_{l=1}^{m} (x_{ik} - \bar{x})^2, \qquad (7.25)$$

где m — число частных совокупностей признака;

3) *средняя внутренних дисперсий*

$$D_r = s^2 = \frac{D_{r_1} + D_{r_2} + \ldots + D_{r_m}}{m}; \qquad (7.26)$$

4) *межгрупповая (внешняя) дисперсия* D_M, представляющая оценку рассеяния групповых средних $\bar{x}_i$ вокруг общей средней $\bar{x}$:

$$D_M = \frac{(\bar{x}_1 - \bar{x})^2 + (\bar{x}_2 - \bar{x})^2 + \ldots + (\bar{x}_m - \bar{x})^2}{m}. \qquad (7.27)$$

Общая дисперсия статистической совокупности S, состоящей из S_1, S_2, ..., S_m частных совокупностей объемом l каждая, равна сумме средней внутренних дисперсий и межгрупповой (внешней) дисперсии, т. е.

$$D = D_r + D_M. \qquad (7.28)$$

Нетрудно заметить, что в том случае, когда все внутренние средние равны общей средней, т. е. когда $\bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_m = \bar{x}$, межгрупповая дисперсия $D_M = 0$, а общая дисперсия равна средней внутренних дисперсий, т. е. $D = D_r$. В остальных случаях общая дисперсия больше, чем средняя внутренних дисперсий на величину, равную величине межгрупповой дисперсии.

Равенство (7.28) имеет место тогда, когда объемы частных совокупностей равны l. Однако, как и правило сложения средних, правило сложения дисперсий легко распространяется на случай, когда объемы частных совокупностей различны и равны соответственно l_1, l_2, ..., l_m. В этом случае общая дисперсия также равна средней внутренних дисперсий плюс межгрупповая дисперсия при условии, что все значения дисперсий берутся взвешенными по объемам l_1, l_2, ..., l_m.

Таким образом, общая дисперсия в этом случае равна

$$D = \frac{l_1 D_{r_1} + l_2 D_{r_2} + \ldots + l_m D_{r_m}}{l_1 + l_2 + \ldots + l_m} + \frac{l_1 (\bar{x}_1 - \bar{x})^2 + l_2 (\bar{x}_2 - \bar{x})^2 + \ldots + l_m (\bar{x}_m - \bar{x})^2}{l_1 + l_2 + \ldots + l_m} =$$

$$= \frac{\sum_{k=1}^{m} l_k D_{r_k}}{\sum_{k=1}^{m} l_k} + \frac{\sum_{k=1}^{m} l_k (\bar{x}_k - \bar{x})^2}{\sum_{k=1}^{m} l_k},$$

или

$$D = D_r + D_M.$$

13. Длина словоупотребления как статистико-стилистический параметр. Величины средней арифметической и дисперсий используются при выявлении статистических характеристик (параметров) стилей [34, с.71—73, 150—164, 178—192, 231 и др.].

Рассмотрим в этом плане средние длины словоформ и рассеяние этих длин в казахской прозе. Возьмем по десять словоупотреблений из четырех разновидностей современного казахского литературного языка — публицистики, художественной прозы (беллетристики), драматургии и научного повествования — и запишем распределение длин этих словоупотреблений в каждой из указанных разновидностей [длины отдельных словоупотреблений приведены в столбцах (3)—(12) табл. 7.23].

Таблица 7.23

Средние длины словоформ и их рассеяние в четырех разновидностях современной казахской прозы

Разновидности литературного языка	Номера в выборке l	1	2	3	4	5	6	7	8	9	10	$\bar{x}_i$	D_{r_i}
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Публицистика	1	12	13	9	6	11	6	8	8	3	6	8,2	8,76
Беллетристика	2	3	6	7	8	5	6	3	3	3	7	5,1	3,49
Драматургия	3	5	5	5	3	2	6	2	2	3	6	3,9	2,49
Научная проза	4	11	5	7	6	17	11	4	13	4	10	8,8	16,76

Пользуясь формулой (7.4), вычисляем среднюю арифметическую

$$\bar{x}_1 = \frac{3 + 6 \cdot 3 + 8 \cdot 2 + 9 + 11 + 12 + 13}{10} = 8{,}2.$$

Остальные значения средних приведены в столбце (13) табл. 7.23.

По формуле (7.24) находим внутреннюю дисперсию:

$$D_{r_1} = \frac{1}{10}[(12-8{,}2)^2+(13-8{,}2)^2+(9-8{,}2)^2+(6-8{,}2)^2+$$
$$+(11-8{,}2)^2+(6-8{,}2)^2+(8-8{,}2)^2+(8-8{,}2)^2+(3-8{,}2)^2+$$
$$+(6-8{,}2)^2] = \frac{1}{10}\cdot 87{,}6 = 8{,}76.$$

Значения D_{r_2}, D_{r_3}, D_{r_4} помещены в столбце (14) табл. 7.23. С помощью формулы (7.26) находим значение средней внутренних дисперсий, которое равно

$$D_r = \frac{8{,}76+3{,}49+2{,}49+16{,}76}{4} = \frac{31{,}50}{4} = 7{,}875.$$

Значение межгрупповой дисперсии D_M, вычисляемое по формуле (7.27), составляет

$$D_M = \frac{1}{4}[(8{,}2-6{,}5)^2+(5{,}1-6{,}5)^2+(3{,}9-6{,}5)^2+(8{,}8-6{,}5)^2] =$$
$$= \frac{1}{4}(1{,}7^2+1{,}4^2+2{,}6^2+2{,}3^2) = \frac{16{,}90}{4} = 4{,}225.$$

Следовательно, значение общей дисперсии согласно формуле (7.28) равно

$$D = s^2 = 7{,}875 + 4{,}255 = 12{,}1,$$

а среднеквадратическое отклонение и коэффициент вариации для общей совокупности соответственно составляют

$$s = \sqrt{12{,}1} = 3{,}48; \quad V = \frac{3{,}48}{(1/4)(8{,}2+5{,}1+3{,}9+8{,}8)}100\% = 53{,}5\%.$$

Значительное рассеяние средних арифметических по жанрам казахской прозы говорит о том, что средняя длина словоупотребления выступает в каждом жанре в качестве стилистико-статистического параметра.

14. Средняя арифметическая и дисперсия в совокупностях с качественным признаком. Статистическая однородность текста. Рассмотренные до сих пор статистические характеристики относились к количественным лингвистическим признакам. Для того чтобы применить понятия средней арифметической и дисперсии к совокупности, которая характеризуется качественным признаком, необходимо эту последнюю преобразовать в совокупность с количественными признаками.

Проще всего осуществить такое преобразование в том случае, когда совокупность содержит объекты, характеризующиеся некоторым признаком, и объекты, не имеющие этого признака. Такую совокупность мы будем называть *статистической совокупностью о альтернативным признаком.*

Отметим наличие признака единицей, а его отсутствие — нулем. Теперь данную качественную совокупность можно представить в виде вариационного ряда, показанного в табл. 7.24.

Таблица 7.24

x_i	$x_1=1$	$x_2=0$
n_i	F	$N-F$

Объем этой совокупности есть величина N, представляющая собой сумму величин F (числа единиц) и $N-F$ (числа нулей). Относительная частота (доля) признака равна $f = \frac{F}{N}$, а доля его отсутствия $\frac{N-F}{N} = 1-f$.

Найдем теперь среднюю арифметическую $\bar{x}$ и дисперсию D. Пользуясь формулой (7.4), нетрудно показать, что средняя арифметическая равна частости:

$$\bar{x} = \frac{F\cdot 1+(N-F)0}{N} = \frac{F}{N} = f. \qquad (7.29)$$

На основании соотношения (7.14) находим дисперсию данной совокупности:

$$D = s^2 = \frac{F\left(1-\frac{F}{N}\right)^2+(N-F)\left(0-\frac{F}{N}\right)^2}{N} =$$
$$= \frac{F(N-F)^2+(N-F)F^2}{N^3} = \frac{F(N-F)}{N^2} = \frac{F}{N}\cdot\frac{N-F}{N}. \qquad (7.30)$$

Дисперсия, являясь здесь мерой рассеивания вариации, характеризует совокупность с качественным признаком с точки зрения ее однородности. Например, полная однородность текста относительно признака A имеет место тогда, когда все его словоупотребления обладают признаком A (т. е. $F = N$) или когда этот признак полностью у них отсутствует (т. е. $F = 0$). Одновременно чем ближе значение $D = s^2$ к нулю, тем однороднее статистическая совокупность (в данном случае текст). Чем больше значение D, тем она разнороднее.

Рассмотрим следующий пример. В целях определения статистической однородности английских научно-технических и разговорных текстов с точки зрения использования в них именных форм подверглись сравнению два отрывка текста из указанных стилей по 1000 словоупотреблений в каждом. В научно-техническом тексте обнаружено 300, а в разговорном — 200 именных словоформ.

Вычислим и сравним дисперсии для научно-технического (D_1) и разговорного (D_2) стилей. Здесь $N = 1000$, $F_1 = 300$, $F_2 = 200$. Следовательно,

$$\bar{x}_1 = 300/1000 = 0.3,\ \bar{x}_2 = 200/1000 = 0{,}2,$$

$$D_1 = (300 \cdot 700)/(1000 \cdot 1000) = 0{,}21,$$

$$D_2 = (200 \cdot 800)/(1000 \cdot 1000) = 0{,}16.$$

Сравнив значения D_1 и D_2, убеждаемся, что разговорный текст с точки зрения употребления именных форм более однороден, чем английский научно-технический текст.

Если исследуемая лингвистическая совокупность является не генеральной, а выборочной совокупностью, как это имело место в рассматриваемых примерах, то формулы (7.29) и (7.30) дают лишь выборочные характеристики $\bar{x}$ и D, но не истинные значения средней арифметической и дисперсии, присущие всей генеральной совокупности. В связи с этим снова возникает уже ставившийся нами в гл. 6 вопрос о близости этих выборочных характеристик к характеристикам генеральной совокупности Этот вопрос будет рассмотрен подробно в гл. 8 и 9.

§ 4. Исследование лингвистических вариационных рядов с помощью эмпирических моментов

Средняя арифметическая и дисперсия представляют собой частные случаи эмпирических (статистических) моментов, являющихся характеристиками опытного распределения (вариационного ряда) лингвистических признаков.

Эмпирический момент h-го порядка определяется как средняя арифметическая h-х степеней отклонений вариант признака X от некоторой произвольно взятой постоянной a (так называемого начала отсчета или условного нуля) и выражается равенством

$$n_h(a) = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - a)^h. \qquad (7.31)$$

Эмпирические моменты являются аналогами теоретических моментов, рассмотренных в гл. 6, § 2, п. 8. По аналогии с теоретическими моментами в зависимости от значения произвольной постоянной a различаются четыре вида эмпирических моментов: моменты относительно постоянной a (при условии, что $a \neq 0$, $a \neq \bar{x}$), начальные моменты (при $a = 0$), центральные моменты (при $a = \bar{x}$) и нормированные моменты.

1. Начальные эмпирические моменты для непрерывных и дискретных лингвистических вариационных рядов. Формулы для первых пяти начальных моментов относительно произвольно фиксированной постоянной a имеют вид

$$n_0(a) = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - a)^0 = 1, \qquad (7.32)$$

$$n_1(a) = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - a), \qquad (7.33)$$

$$n_2(a) = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - a)^2, \qquad (7.34)$$

$$n_3(a) = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - a)^3, \qquad (7.35)$$

$$n_4(a) = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - a)^4. \qquad (7.36)$$

В том случае, когда разности $x_i - a$ имеют общий множитель l, вычисление моментов можно упростить, если воспользоваться выражением

$$n'_h = \frac{1}{N}\sum_{i=1}^{k} n_i \left(\frac{x_i - a}{l}\right)^h = \frac{S_h}{N}, \qquad (7.37)$$

где

$$S_h = \sum_{i=1}^{k} n_i \left(\frac{x_i - a}{l}\right)^h, \qquad (7.38)$$

Величины $(x_i - a)/l = x'_i$ выступают в качестве условных (рабочих) вариант случайной величины, а n'_h является рабочим начальным моментом.

Если необходимо получить истинные значения начальных моментов относительно a, то правую часть равенства (7.37) нужно умножить на h-ю степень множителя l. В итоге получаем

$$n_h(a) = \frac{1}{N}\sum_{i=1}^{k} n_i \left(\frac{x_i - a}{l}\right)^h l^h = \frac{S_h l^h}{N}. \qquad (7.39)$$

Этим приемом мы уже пользовались при вычислении средней арифметической (§ 3, п.3) и опытной дисперсии (§ 3, п. 11).

Теперь воспользуемся методом моментов для определения характеристик непрерывного распределения — вариационного ряда длин китайских слогов (см. § 2, п.2).

Полагая $a = 145$, $l = 30$ и пользуясь формулами (7.37) — (7.39), произведем расчеты, как это показано в табл. 7.25. В итоге получаем величины n_1, n_2, n_3, n_4, которые показаны в нижней строке таблицы ($n_0 = 1$).

Расчет начальных моментов из-за своей громоздкости часто сопровождается ошибками в вычислениях. Для проверки расчетов

Таблица 7.25

Интервалы длин слогов в (мс) (1)	Середины интервалов x_i (2)	l_u (3)	$v - {}^{l}x$ (4)	$l_x = \frac{l}{v - {}^{l}x}$ (5)	$\frac{l}{v - {}^{l}x}\, l_u$ (6)	$\left(\frac{l}{v - {}^{l}x}\right)^2 l_u$ (7)	$\left(\frac{l}{n - {}^{l}x}\right)^3 l_u$ (8)	$\left(\frac{l}{v - {}^{l}x}\right)^4 l_u$ (9)	$1 + \frac{l}{v - {}^{l}x}$ (10)	$\left(1 + \frac{l}{v - {}^{l}x}\right)^4 l_u$ (11)
40—70	55	7	—90	—3	—21	63	—189	567	—2	112
70—100	85	9	—60	—2	—18	36	—72	144	—1	9
100—130	115	49	—30	—1	—49	49	—49	49	0	0
130—160	a=145	33	0	0	0	0	0	0	1	33
160—190	175	22	30	1	22	22	22	22	2	352
190—220	205	18	60	2	36	72	144	288	3	1458
220—250	235	7	90	3	21	63	189	567	4	1792
250—280	265	3	120	4	12	48	192	768	5	1875
280—310	295	2	150	5	10	50	250	1250	6	2592
		N=150			S_1=13	S_2=403	S_3=487	S_4=3655		S_4^*=8223
					n_1'=13/150	n_2'=403/150	n_3'=487/150	n_4'= 3655/150		n_4^*=8223/150
					n_1=2,6=$\bar{x}$	n_2=2418	n_3=37 660	n_4=19 737 000		

обычно пользуются формулой, выражающей связь первых пяти начальных моментов для вариант x_i с начальным моментом четвертого порядка для вариант $x_i + 1$. Эта формула имеет вид

$$n_4'^* = n_0' + 4n_1' + 6n_2' + 4n_3' + n_4', \tag{7.40}$$

где $n_4'^*$ — четвертый начальный момент вариант $x_i + 1$, а правая часть равенства включает пять начальных моментов вариант x_i.

Если значение момента $n_4'^*$, вычисленное по формуле (7.40), совпадает с его значением, полученным из выражения (7.36), то это значит, что расчет первых пяти моментов для вариант x_i произведен правильно.

Проверим правильность вычисления рабочих начальных моментов в вариационном ряде длин китайских слогов. Для этого, используя данные столбца (11) табл. 7.25, находим, что $n_4'^* = 8223/150$. Эта величина совпадает с величиной n_4^*, вычисленной с помощью формулы (7.40):

$$n_4'^* = 1 + 4 \cdot \frac{13}{150} + 6 \cdot \frac{403}{150} + 4 \cdot \frac{487}{150} + \frac{3655}{150} = \frac{150 + 52 + 2418 + 1948 + 3655}{150} = \frac{8223}{150}.$$

Таким образом, рабочие начальные моменты для распределения длин китайских слогов определены правильно.

Если вариационный ряд лингвистических признаков содержит нулевую варианту, то целесообразно пользоваться начальными эмпирическими моментами.

Формулы для первых пяти начальных эмпирических моментов имеют вид:

$$n_0 = \frac{1}{N} \sum_{i=1}^{k} n_i x_i^0 = \frac{S_0}{N} = 1, \tag{7.41}$$

$$n_1 = \frac{1}{N} \sum_{i=1}^{k} n_i x_i = \frac{S_1}{N} = \bar{x}, \tag{7.42}$$

(*средняя арифметическая*),

$$n_2 = \frac{1}{N} \sum_{i=1}^{k} n_i x_i^2 = \frac{S_2}{N} = \bar{x}^2, \tag{7.43}$$

$$n_3 = \frac{1}{N} \sum_{i=1}^{k} n_i x_i^3 = \frac{S_3}{N} = \bar{x}^3; \tag{7.44}$$

$$n_4 = \frac{1}{N} \sum_{i=1}^{k} n_i x_i^4 = \frac{S_4}{N} = \bar{x}^4; \tag{7.45}$$

причем

$$S_h = \sum_{i=1}^{k} n_i x_i^h. \tag{7.46}$$

Воспользуемся приведенными выше сведениями при решении задачи автоматического устранения многозначности слова в тексте. Это устранение осуществляется путем поиска в самом слове или в его окружении некоторого формального признака — *индикатора*. Этот индикатор в комбинации с другими индикаторами или же сам по себе помогает машине выбрать из автоматического словаря подходящее для данного контекста значение слова. Индикаторами могут быть как морфемы, так и отдельные слова. Для унификации поиска та позиция в предложении, которую занимает многозначное слово, считается нулевой, позиции слева от контрольного слова помечаются целыми отрицательными числами, а позиции справа — положительными.

Для оптимизации поиска индикатора бывает важно иметь распределение частот индикаторов в различных позициях и получить характеристики этого распределения.

Так, например, распределение 100 индикаторов в контрольном отрывке немецкого научно-технического текста характеризуется величинами, показанными в столбцах (1) и (2) табл. 7.26.

Таблица 7.26

Позиции индикаторов x_i	Частота индикаторов n_i	$n_i x_i$	$n_i x_i^2$	$n_i x_i^3$	$n_i x_i^4$	x_i+1	$n_i (x_i+1)^4$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
—2	10	—20	40	—80	160	—1	10
—1	30	—30	30	—30	30	0	0
0	30	0	0	0	0	1	30
1	20	20	20	20	20	2	320
2	10	20	40	80	160	3	810
	$N=100$	$S_1=-10$	$S_2=130$	$S_3=-10$	$S_4=370$		$S_4^*=1170$
		$n_1'=-0,1$	$n_2'=1,3$	$n_3'=-0,1$	$n_4'=3,7$		$n_4^*=11,7$

Данный вариационный ряд содержит нулевую варианту, в качестве которой выступает позиция многозначного слова, поэтому постоянное число $a = 0$. Отсюда следует, что характеристиками рассматриваемого ряда служат начальные моменты, расчет которых следует вести по формулам (7.41)—(7.46), как это показано в табл. 7.26.

Теперь проверим правильность расчетов при определении первых начальных моментов в распределении частот индикаторов устранения многозначности в немецком научно-техническом тексте.

Для этого, пользуясь данными табл. 7.26, находим четвертый начальный момент для вариант $x_i + 1$:

$$n_4'^* = 1170/100 = 11,7.$$

Тот же результат получаем с помощью формулы (7.40):

$$n_4'^* = 1 + 4(-0,1) + 6 \cdot 1,3 + 4(-0,1) + 3,7 = 11,7.$$

Таким образом, вычисления начальных моментов произведены правильно.

Тот факт, что первый начальный момент, совпадающий со средней арифметической, имеет близкое к нулю отрицательное значение ($n_1' = \bar{x} = -0,1$), говорит о том, что снимающий полисемию индикатор следует искать в непосредственном левом окружении многозначного слова.

2. Центральные и нормированные эмпирические моменты. Начальные моменты не только используются в лингвистике при определении средней арифметической, но и применяются в основном как вспомогательные величины при вычислении центральных эмпирических моментов, которые дают основную характеристику рассеяния случайных величин.

Формулы для вычисления эмпирических центральных моментов аналогичны формулам теоретических центральных моментов (см. гл. 6, § 2, п.8):

$$m_0 = 1,$$

$$m_1 = 0,$$

$$m_2 = s^2 = n_2 - n_1 \quad (\textit{опытная дисперсия}), \tag{7.47}$$

$$m_3 = n_3 - 3n_2n_1 + 2n_1^3, \tag{7.48}$$

$$m_4 = n_4 - 4n_3n_1 + 6n_2n_1^2 - 3n_1^4. \tag{7.49}$$

Если квадратный корень из второго центрального момента $\sqrt{m_2} = s$ принять за стандарт, то отношение центрального момента h-го порядка к h-й степени стандарта даст безразмерный *нормированный эмпирический момент*

$$r_h = m_h/(\sqrt{m_2})^h = m_h/s^h, \tag{7.50}$$

аналогичный теоретическому нормированному моменту ρ_h.

Для первых четырех нормированных эмпирических моментов имеем:

$$r_1 = 0,$$

$$r_2 = 1,$$

$$r_3 = m_3/(\sqrt{m_2})^3 = m_3/s^3, \tag{7.51}$$

$$r_4 = m_4/m_2^2 = m_4/s^4. \tag{7.52}$$

Третий и четвертый нормированные моменты подобно их теоретическим аналогам ρ_3 и ρ_4 служат соответственно для количественной оценки асимметрии (скошенности) и крутости (островершинности или плосковершинности) эмпирического распределения.

Если r_3, а также все нечетные центральные моменты m_1, m_3, m_5, ... равны нулю, то эмпирическое распределение является симметрич-

ным. При $r_3 > 0$ ряд характеризуется правосторонней (положительной), а при $r_3 < 0$ — левосторонней (отрицательной) скошенностью (ср. гл. 6, § 2, п. 8).

При $r_4 > 3$ эмпирическое распределение характеризуется острой вершиной; плосковершинный или двухвершинный ряд дает $r_4 < 3$. Для измерения крутизны распределения используется также величина опытного эксцесса

$$E_э = r_4 - 3, \qquad (7.53)$$

аналогичная теоретическому эксцессу (куртозису).

3. Использование эмпирических моментов для сравнения лингвостатистических вариационных рядов с теоретическими распределениями. Часто характер эмпирического распределения лингвистиче-

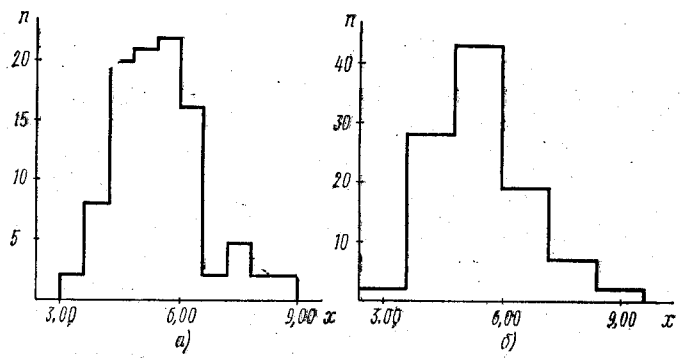


Рис. 56

ских единиц оценивается непосредственно по виду графика, полигона или гистограммы. При этом не учитывается тот факт, что графический вид распределения во многом определяется выбором границ и ширины интервала. Например, на рис. 56 показаны гистограммы одного и того же распределения средних длин словоформ в языках мира (исходное распределение этих длин словоформ приведено в табл. 7.27). Первая гистограмма (а) построена на основании данных табл. 7.28, использующей узкий интервал. Вторая же (б) опирается на ряд, который использует вдвое более широкий интервал и дает смещение конца распределения вправо (табл. 7.29). При этом может показаться, что оба распределения существенно отличаются друг от друга, хотя в действительности мы имеем дело с одним и тем же эмпирическим распределением. Аналогичным образом можно получить разные гистограммы распределения длин китайских слогов (см. табл. 7.9 и 7.25) или индикаторов снятия многозначности в немецких текстах (табл. 7.26).

Поэтому, чтобы не впасть в ошибку, предварительную оценку характера эмпирического распределения следует производить не визуально, а опираясь на количественные характеристики вариационного ряда. Для описания лингвистических вариационных ря-

Таблица 7.27

Средние длины словоформ в некоторых языках мира

№	Языки	x	№	Языки	x	№	Языки	x
1	Русский	4,701	35	Бретонский	3,904	66	Алтайский	5,750
2	Украинский	5,156	36	Ирландский	3,673	67	Гагаузский	5,069
3	Белорусский	4,742	37	Уэльский	4,450	68	Карачаевский	5,415
4	Болгарский	4,186	38	Курдский	3,811	69	Балкарский	6,023
5	Македонский	5,037	39	Таджикский	6,205	70	Туркменский	5,753
6	Сербохорватский	4,654	40	Осетинский	4,784	71	Узбекский	7,471
7	Словенский	4,423	41	Бенгали	5,360	72	Уйгурский	5,556
8	Чешский	4,145	42	Албанский	3,776	73	Якутский	5,764
9	Словацкий	4,280	43	Армянский	6,512	74	Тувинский	8,776
10	Польский	6,531	44	Финский	7,382	75	Хакасский	6,340
11	Кашубский	4,570	45	Эстонский	7,754	76	Бурятский	5,952
12	Лужицкий	4,940	46	Коми	4,766	77	Калмыцкий	4,614
13	Литовский	5,775	47	Коми-Пермяцкий	5,493	78	Эвенкийский	5,778
14	Латышский	5,786	48	Удмуртский	5,299	79	Эвенский	6,140
15	Шведский	4,527	49	Марийский (луговой)	4,710	80	Нанайский	6,571
16	Датский	4,781	50	Марийский (горный)	5,117	81	Дунганский	4,613
17	Норвежский	4,222	51	Мордовско-эрзянский	5,356	82	Чукотский	8,441
18	Исландский	5,000	52	Мордовско-мокшанский	5,477	83	Нивхский	8,340
19	Фарерский	5,589	53	Ненецкий	6,217	84	Эскимосский	8,296
20	Немецкий	5,448	54	Селькупский	5,326	85	Кетский	6,109
21	Голландский	4,984	55	Венгерский	4,841	86	Абхазский	6,406
22	Фламандский	4,463	56	Мансийский	5,753	87	Абазинский	6,016
23	Фризский	4,767	57	Хантыйский	4,126	88	Адыгейский	7,565
24	Английский	3,042	58	Азербайджанский	6,308	89	Кабардино-черкесский	7,156
25	Французский	4,855	59	Татарский	5,929	90	Чеченский	3,424
26	Провансальский	4,654	60	Башкирский	5,507	91	Ингушский	5,022
27	Итальянский	4,500	61	Кумыкский	5,474	92	Аварский	7,300
28	Испанский	4,953	62	Каракалпакский	5,589	93	Даргинский	7,194
29	Каталанский	5,638	63	Ногайский	5,584	94	Табасаранский	6,500
30	Португальский	4,829	64	Казахский	5,216	95	Лакский	5,520
31	Румынский	6,164	65	Киргизский	5,809	96	Баскский	5,800
32	Молдавский	4,968				97	Хауса	4,720
33	Ретороманский	3,946				98	Суахили	5,209
34	Латинский	5,072				99	Вьетнамский	4,979
						100	Индонезийский	6,253

Таблица 7.28

Интервалы		Интервалы	
2,40—3,00	0	6,00—6,60	16
3,00—3,60	2	6,60—7,20	2
3,60—4,20	8	7,20—7,80	5
4,20—4,80	20	7,80—8,40	2
4,80—5,40	21	8,40—9,00	2
5,40—6,00	22		
			$N=100$

Таблица 7.29

Интервалы		Интервалы	
2,40—3,60	2	6,00—7,20	18
3,60—4,80	28	7,20—8,40	7
4,80—6,00	43	8,40—9,60	2
			$N=100$

дов используются следующие четыре эмпирических момента, аналогичные тем четырем моментам, которые характеризовали теоретические распределения (см. гл. 6, § 2, п. 8):

1) первый начальный момент $n_1 = \bar{x}$ (средняя арифметическая),
2) второй центральный момент $m_2 = s^2$ (опытная дисперсия),
3) третий нормированный момент $r_3 = m_3/s^3$ (коэффициент скошенности ряда),
4) четвертый нормированный момент $r_4 = m_4/s^4$ (коэффициент крутости ряда).

Пользуясь численными значениями отдельных моментов, а также сравнивая эти значения между собой, можно выдвигать гипотезы о соответствии данного вариационного ряда тому или иному распределению. Выше было показано, что нормальное распределение имеет, в частности, параметры $\rho_3 = 0$, $\rho_4 = 3$. Поэтому если третий эмпирический нормированный момент r_3 вариационного ряда не сильно отличается от нуля, что указывает на симметрию ряда, а значение четвертого нормированного момента r_4 близко к трем, что свидетельствует о средней крутости ряда, то можно выдвинуть гипотезу о нормальности рассматриваемого нами распределения (о методах проверки этой гипотезы см. в гл. 9).

Если же величина r_3 заметно отличается от нуля, а $r_4 \gg 3$ или $r_4 \ll 3$, то предположение о нормальности лингвистического вариационного ряда должно быть сразу же отвергнуто. Такие оценки широко использовались, например, при определении нормальности эмпирических распределений слов и суффиксов в различных подъязыках и стилях современного латышского языка (31, с.12—13). Следует иметь в виду, что здесь не указываются те границы, которых

могут достигать отклонения эмпирических моментов от их теоретических аналогов, т. е. границы, в рамках которых лингвистический вариационный ряд сохраняет свою принадлежность к нормальному распределению.

Эти границы могут быть определены исходя из следующих соображений. Будем рассматривать значение третьего нормированного момента r_3 как показатель суммарного смещения кривой эмпирического распределения наблюдаемых погрешностей от центра кривой нормального распределения. Это смещение можно оценить с помощью среднего квадратического отклонения

$$\sigma(r_3) = \sqrt{\frac{6(N-1)}{(N+1)(N+3)}}, \tag{7.54}$$

где N — объем выборки. Если N велико, то вместо (7.54) можно пользоваться приближенной оценкой

$$\sigma(r_3) \approx \sqrt{6/N}. \tag{7.55}$$

Мерой согласованности эмпирического распределения с нормальным законом служит отношение

$$|r_3|/\sigma(r_3) = \alpha. \tag{7.56}$$

Если $\alpha < 3$, то смещение эмпирической кривой относительно центра нормального распределения можно считать несущественным.

Для оценки расхождения эмпирического распределения и теоретического нормального распределения используется среднее квадратическое отклонение

$$\sigma(E) = \sqrt{\frac{24N(N-2)(N-3)}{(N-1)^2(N+3)(N+5)}}, \tag{7.57}$$

причем при больших N соотношение (7.57) заменяется приближенным равенством

$$\sigma(E) \approx 2\sqrt{6/N} = 2\sigma(r_3). \tag{7.58}$$

Соответственно мерой согласованности эмпирического и теоретического нормального распределений служит отношение эксцесса к среднему квадратическому $\sigma(E)$:

$$E_э/\sigma(E) = \beta. \tag{7.59}$$

Если $\beta < 3$, то отличие вариационного ряда от нормального распределения несущественно.

Итак, если для рассматриваемого лингвистического вариационного ряда $\alpha < 3$ и $\beta < 3$, то можно предполагать, что этот ряд имеет нормальное распределение. Подробнее о применении этой процедуры см. ниже.

Значения опытных моментов могут быть использованы также для сравнения вариационного ряда и теоретического распределения Пуассона. Поскольку для распределения Пуассона имеет место равенство теоретических моментов $\nu_1 = \mu_2 = \mu_3 = \lambda$ (см. гл. 6,

§ 3, п. 2), то лингвистический вариационный ряд, в котором значения эмпирических моментов n_1, m_2 и m_3 мало отличаются друг от друга по величине, может быть гипотетически отнесен к распределению Пуассона.

В работе Т. А. Якубайтис [31, с. 7—46, 59—66] вычислены эмпирические моменты n_1, m_2, m_3 для распределения 120 слов, 6 суффиксов и класса частиц в различных подъязыках и функциональных стилях латышского языка. Обнаружилось, что по мере уменьшения частоты слов численные значения эмпирических моментов n_1, m_2, m_3 сближаются. Отсюда можно сделать вывод, что распределение некоторых редких слов ($F > 70$) в латышском языке соответствует закону Пуассона.

4. Распределение средних длин словоупотреблений в языках мира. В § 3, п. 13 было показано, что средняя длина словоупотребления может выступать в качестве статистического параметра стиля. В связи с этим возникает вопрос: можно ли считать среднюю длину словоупотребления типологической характеристикой конкретного языка [8]; [51, с. 178—181]? Статистический анализ распределения средних длин словоупотреблений в языках мира помогает решению этого вопроса. Действительно, если вариационный ряд средних длин хорошо аппроксимируется нормальной кривой, то можно предполагать, что средние длины словоупотреблений разных языков равномерно группируются вокруг некоторой средней, задаваемой возможностями оперативной памяти человека. Отклонения от этой средней в каждом конкретном языке следует рассматривать как результат случайных воздействий, по всей вероятности, не связанных с типологией языка.

Заметное отклонение вариационного ряда от нормальной кривой можно рассматривать как указание на то, что распределение средних длин словоформ определяется не только случайными, но и детерминированными процессами. В этом случае естественно предположить, что расхождения в средних длинах словоупотреблений связаны с типологическими особенностями конкретных языков.

Чтобы решить, какое из этих двух предположений считать более правдоподобным, обратимся к статистическому анализу распределения средних длин словоупотреблений для ста языков мира, применяющих буквенную графику*. Для этого используем приведенные выше табл. 7.27 и 7.28.

Статистический анализ ряда, проведенный строго по схеме, которая была изложена в п. 1—3, показан в табл. 7.30. Начнем с того, что определим центральное значение $\hat{x}_i$ каждого интервала [столбец (2)]. Затем вычислим начальные моменты относительно a, в качестве которого возьмем среднее значение центрального интервала ($a = 5{,}70$). Используя общий для разностей $\hat{x}_i - a$ множитель $l = 0{,}6$

* Расчет произведен по разговорно-беллетристическим текстам этих языков, а также по отрывкам, приведенным в книге [12]. Если учитывать научно-техническую и публицистическую прозу, то значения средних длин будут несколько выше.

Таблица 7.30

Интервалы	(1)	$\hat{x}_i$ (2)	u_i (3)	$\hat{x}_i - a$ (4)	$\frac{1}{v - \hat{x}_i}$ (5)	$u_i \left(\frac{1}{v-\hat{x}_i}\right)$ (6)	$u_i \left(\frac{1}{v-\hat{x}_i}\right)^2$ (7)	$u_i \left(\frac{1}{v-\hat{x}_i}\right)^3$ (8)	$u_i \left(\frac{1}{v-\hat{x}_i}\right)^4$ (9)	$1 + \frac{1}{v-\hat{x}_i}$ (10)	$\left(1 + \frac{1}{v-\hat{x}_i}\right)^4$ (11)	$u_i \left(1 + \frac{1}{v-\hat{x}_i}\right)^4$ (12)	$\hat{x}_i u_i$ (13)	$\hat{x}_i - \bar{x}$ (14)	$(\hat{x}_i - \bar{x})^2$ (15)	$u_i (\hat{x}_i - \bar{x})^2$ (16)	$(\hat{x}_i - \bar{x})^3 u_i$ (17)	$(\hat{x}_i - \bar{x})^4 u_i$ (18)
2,40—3,00		2,70	0	—3,00	—5	0	0	0	0	—4	256	0	0	—	—	0	0	0
3,00—3,60		3,30	2	—2,40	—4	— 8	32	—128	512	—3	81	162	6,6	—2,16	4,6656	9,3312	—20,1554	43,5357
3,60—4,20		3,90	8	—1,80	—3	—24	72	—216	648	—2	16	128	31,2	—1,56	2,4336	19,4688	—30,3713	47,3792
4,20—4,80		4,50	20	—1,20	—2	—40	80	—160	320	—1	1	20	90	—0,96	0,9216	18,4320	—17,6947	16,9869
4,80—5,40		5,10	21	—0,60	—1	—21	21	— 21	21	0	0	0	107,1	—0,36	0,1296	2,7216	— 0,9798	0,3527
5,40—6,00	a	=5,70	22	0	0	0	0	0	0	1	1	22	125,4	0,24	0,0576	1,2672	0,3041	0,0730
6,00—6,60		6,30	16	0,60	1	16	16	16	16	2	16	256	100,8	0,84	0,7056	11,2896	9,4833	7,9660
6,60—7,20		6,90	2	1,20	2	4	8	16	32	3	81	162	13,8	1,44	2,0736	4,1472	5,9720	8,6000
7,20—7,80		7,50	5	1,80	3	15	45	135	405	4	256	1280	37,5	2,04	4,1616	20,8080	42,4483	86,5945
7,80—8,40		8,10	2	2,40	4	8	32	128	512	5	625	1250	16,2	2,64	6,9696	13,9392	36,7995	97,1507
8,40—9,00		8,70	2	3,09	5	10	50	250	1250	6	1296	2592	17,4	3,24	10,4976	20,9952	68,0244	220,3991
						$S_1 = -40$	$S_2 = 356$	$S_3 = 20$	$S_4 = 3716$			$S_4^* = 5872$	546					529,0378
						$n_1' = -0{,}24$	$n_2' = 3{,}56$	$n_3' = 0{,}20$	$n_4' = 37{,}16$			$n_4'^* = 58{,}72$	$\bar{x} = 5{,}46$			$m_2 = 1{,}224$	$m_3 = 0{,}9384$	$m_4 = 5{,}2903$
						$n_1 = -0{,}24$	$n_2 = 1{,}2816$	$n_3 = 0{,}0432$	$n_4 = 4{,}8159$									

[столбцы (4) и (5)], определим сначала рабочие начальные моменты. Для этого, разделив каждую из сумм столбцов (6) —(9) на $N = 100$, получим:

$$n'_0 = 1,\ n'_1 = -0,4,\ n'_2 = 3,56,\ n'_3 = 0,2,\ n'_4 = 37,16.$$

Для проверки правильности вычислений определим четвертый момент для условных вариант $\frac{\hat{x}-a}{l} + 1$ с помощью равенства (7.40):

$$n'^{*}_4 = 1 + 4\cdot(-0,4) + 6\cdot 3,56 + 4\cdot 0,2 + 37,16 = 58\ 72.$$

Этот результат совпадает со значением n'^{*}_4, полученным путем непосредственного вычисления по формуле (7.36) [ход расчетов показан в столбцах (10)—(12) табл. 7.30]. Таким образом, расчет рабочих начальных моментов получен правильно.

Теперь можно перейти к вычислению начальных моментов относительно $a = 5,7$. Для этого каждое из значений n'_h нужно умножить на $l^h = 0,6^h$. Тогда получим:

$$n_0 = 1\cdot 0,6^0 = 1;\ n_1 = -0,4\cdot 0,6^1 = -0,24;\ n_2 = 3,56\cdot 0,6^2 = 1,2816;\ n_3 = 0,2\cdot 0,6^3 = 0,0432;\ n_4 = 37,16\cdot 0,6^4 = 4,8159.$$

Средняя арифметическая ряда (т. е. средняя длина словоупотребления) в языках мира равна

$$\bar{x} = n_1\ (a) + a = -0,24 + 5,7 = 5,46.$$

Затем, используя начальные моменты относительно a и формулы (7.32)—(7.36), вычислим центральные моменты:

$$m_0 = 1;$$

$$m_1 = 0;$$

$$m_2 = s^2 = 1,2816 - (-0,24)^2 = 1,2816 - 0,0576 = 1,224;$$

$$m_3 = 0,0432 - 3\cdot 1,2816\cdot(-0,24) + 2(-0,24)^3 = 0,0432 + 0,9228 - 0,0276 = 0,9384;$$

$$m_4 = 4,8159 - 4\cdot 0,0432\cdot(-0,24) + 6\cdot 1,2816\cdot(-0,24)^2 - 3\cdot(-0,24)^4 = 4,8159 + 0,0415 + 0,4429 - 0,0100 = 5,2903.$$

Центральные моменты можно получить и с помощью формул (7.47)—(7.49). Для этого каждую сумму из столбцов (16)—(18) разделим на $N = 100$. Нетрудно заметить, что полученные результаты совпадают до третьего знака после запятой со значениями центральных моментов, полученных по упрощенной процедуре. Таким образом, расчет центральных моментов произведен правильно.

Теперь, используя значения центральных моментов, переходим с помощью выражений (7.54)—(7.59) к непосредственной проверке

нормальности распределения средних длин словоупотреблений в языках мира. Имеем:

$$r_3 = 0,9384/(\sqrt{1,224})^3 = 0,693;$$

$$E_3 = 5,2903/1,224^2 - 3 = 3,527 - 3 = 0,527.$$

Далее получаем

$$\sigma(r_3) = \sqrt{(6\cdot 99)/(101\cdot 104)} = \sqrt{0,05655} = 0,238,$$

или по приближенной оценке (7.55),

$$\sigma(r_3) \approx \sqrt{6/100} \approx 0,245,$$

а также

$$\sigma(E) = \sqrt{(2400\cdot 98\cdot 97)/(99^2\cdot 103\cdot 105)} = \sqrt{0,215244} = 0,464,$$

или по оценке (7.58),

$$\sigma\ (E) \approx 0,490.$$

Поскольку отношения

$$\alpha = 0,693/0,238 = 2,91,\ \alpha \approx 0,693/0,245 \approx 2,83,$$

$$\beta = 0,527/0,464 = 1,14,\ \beta \approx 0,527/0,490 \approx 1,08$$

меньше трех, то можно предположить, что распределение средних длин словоформ в языках мира подчиняется нормальному закону. Отсюда следует, что образование словоформ во всех языках мира ориентировано на некоторый эталон, длина которого определена возможностями оперативной памяти человека.

На этом мы заканчиваем описание важнейших приемов формирования вариационных рядов и методики вычисления их характеристик. Каждый вариационный ряд обобщает результаты статистического эксперимента над текстом ограниченной длины, который выступает в роли частной выборки, взятой из некоторой генеральной совокупности (язык в целом, стиль, подъязык, язык писателя). Обычно выборочная текстовая совокупность интересует лингвиста лишь постольку, поскольку она может рассматриваться в качестве модели, отражающей вероятностные свойства исследуемой генеральной совокупности текстов и стоящей за ней нормы языка или его разновидности.

ГЛАВА 8

СТАТИСТИЧЕСКАЯ МОДЕЛЬ ТЕКСТА И ВЕРОЯТНОСТНЫЕ ХАРАКТЕРИСТИКИ НОРМЫ ЯЗЫКА

Переход от статистической модели выборки текста к вероятностным характеристикам нормы языка связан с решением трех задач.

Во-первых, по характеристикам θ^* вариационного ряда необходимо численно оценить скрытые от прямого наблюдения параметры θ соответствующего распределения генеральной совокупности, т. е. параметры, выступающие в качестве вероятностных характеристик нормы языка и его разновидностей.

Во-вторых, по данным вариационного ряда следует оценить характер генерального распределения.

В-третьих, имея в своем распоряжении численные оценки параметров генерального распределения, а также зная характер этого распределения, необходимо решить важнейшую технологическую задачу лингвистического исследования, состоящую в определении того, какой объем обследуемого текста даст достаточно надежные лингвистические результаты.

§ 1. Точечная оценка параметров генеральной лингвистической совокупности

1. Понятие точечной оценки. Каждый параметр θ^* вариационного ряда, вычисленный на основе ограниченного числа опытов, всегда содержит элемент случайности. Поэтому его значение нельзя отождествлять со значением параметра θ, характеризующего распределение исследуемого лингвистического явления в генеральной совокупности. Величину θ^* следует рассматривать лишь как точечную оценку значения θ.

Что же представляет собой эта оценка θ^* с вероятностно-статистической точки зрения? Чтобы ответить на этот вопрос, проведем следующий мысленный эксперимент.

Предположим, что из генеральной совокупности текстов извлекаются 1-я, 2-я, ..., N-я выборки одного и того же объекта, в каждой из которых наблюдаемая случайная величина X с параметром θ^* принимает значения $x_1, x_2, \ldots, x_N$. Все эти значения имеют одинаковые распределения, идентичные распределению величины X. Каждое распределение соответственно характеризуется параметрами $\theta_1^*, \theta_2^*, \ldots, \theta_N^*$, которые являются значениями параметра θ^*. Этот последний представляет собой функцию значений $x_1, x_2, \ldots, x_N$ и является случайной величиной (ведь каждая новая i-я выборка дает новое значение θ^* нашей оценки).

Так как точечная оценка θ^* является случайной величиной, то она может давать разные отклонения от значения θ, и следовательно, давать разную по своей «доброкачественности» оценку параметра θ. Наша задача состоит в том, чтобы найти критерии, позволяющие выбирать из параметров $\theta_\alpha^*, \theta_\beta^*, \ldots, \theta_\lambda^*$ исследованной частной выборки такую величину θ_i^*, которая давала бы наименьшее отклонение от значения θ и выступала бы поэтому в качестве наилучшей оценки.

Действительно, мы пока не знаем, что является наилучшей оценкой математического ожидания $M(X) = \mu$ в генеральной совокупности: средняя арифметическая $\bar{x}$, медиана Me или мода Mo. Аналогичным образом мы не можем пока сказать, что является лучшей оценкой дисперсии $D(X) = \sigma^2$: выборочная дисперсия s^2, размах R или среднее абсолютное отклонение $|x_i - \bar{x}|$.

Чтобы получить критерии выбора наилучших оценок параметров распределений в генеральных совокупностях, необходимо, во-первых, определить те требования к оценкам θ^*, которые давали бы хорошее приближение этих оценок к параметру θ, во-вторых, определить методы нахождения оценок; в-третьих, выяснить возможности использования этих оценок для получения надежных выводов относительно значений параметров генеральной лингвистической совокупности. Хорошее приближение характеристики θ^* к теоретическому параметру θ должно отвечать требованию с о с т о я т е л ь н о с т и, сущность которого заключается в следующем. Идеальной оценкой для параметра θ была бы такая величина θ^*, которая в каждом своем выборочном значении $\theta_1, \theta_2, \ldots, \theta_N^*$ в точности совпала бы с искомым параметром θ. Разумеется, что на практике такое положение недостижимо. К идеальному положению, при котором $\theta^* = \theta$, можно было бы приблизиться, постепенно усредняя значения $\theta_1^*, \theta_2^*, \ldots, \theta_N^*$ при условии, что $N \to \infty$, как это предусмотрено законом больших чисел (см. гл. 6, § 4).

В этом случае согласно закону больших чисел имела бы место сходимость по вероятности величин θ^* к θ, описываемая равенством

$$\lim_{N\to\infty} P\{|\theta^* - \theta| < \varepsilon\} = 1, \qquad (8.1)$$

в котором ε — сколь угодно малая величина.

Итак, оценка неизвестного параметра θ в генеральной лингвистической совокупности отвечает требованию состоятельности, когда она подчиняется закону больших чисел.

Нетрудно заметить, что требование состоятельности имеет два практических недостатка. Во-первых, оно не всегда обнаруживается в условиях малых выборок, с которыми имеет дело языковед; во-вторых, при одной и той же функции распределения генеральной совокупности для данного параметра θ можно найти бесконечное множество состоятельных оценок, сходящихся по вероятности к θ. Поэтому требование состоятельности дополняется еще двумя требованиями: н е с м е щ е н н о с т и и э ф ф е к т и в н о с т и. *Несмещенной оценкой* называется такое приближение, при котором математическое ожидание θ^* равно параметру θ генеральной совокупности, т. е.

$$M(\theta^*) = \theta.$$

Иными словами, несмещенность оценки означает равенство по абсолютной величине взвешенных сумм отклонений значений θ_1^*, θ_2^*, ..., θ_N^* от центра, в качестве которого выступает параметр θ. Вместе с тем она показывает на отсутствие систематической ошибки, постоянно смещающей указанные значения в одну сторону от этого центра.

2. Несмещенные и эффективные точечные оценки математического ожидания и дисперсии. В работах по математической статистике [10, с. 314—317] доказывается, что средняя арифметическая $\bar{x}$ выборочной совокупности является несмещенной оценкой математического ожидания случайной величины генеральной совокупности, т. е.

$$M(\bar{x}) = M(X),$$

и что относительная частота f есть несмещенная оценка вероятности p. Одновременно выясняется, что в малых выборках ($N \leqslant 50$) опытная дисперсия $D_x = s^2$ не может служить несмещенной оценкой теоретической дисперсии $D(X) = \sigma^2$. Несмещенной оценкой теоретической дисперсии является величина

$$\hat{s}^2 = \frac{N}{N-1} s^2. \qquad (8.2)$$

При больших значениях N величиной смещения можно пренебречь и рассматривать s^2 в качестве несмещенной оценки $D(X) = \sigma^2$.

Возможные значения θ могут не давать смещения, однако одновременно могут быть сильно рассеяны вокруг $M(\theta^*)$, что заметно ухудшает оценку параметра θ. Поскольку наилучшей мерой среди рассматриваемых в этой книге мер рассеяния является дисперсия, целесообразно оценивать *эффективность* оценки θ_i^* параметра θ с помощью величины

$$D(\theta_i^*) = \sigma^2(\theta_i^*).$$

Таким образом, если имеется несколько состоятельных несмещенных оценок параметра θ, то наиболее эффективной служит та, которая имеет наименьшую дисперсию и наиболее тесную концентрацию значений θ^* вокруг $M(\theta^*) = \theta$.

В качестве примера рассмотрим две оценки математического ожидания употребления служебных слов, дающих обычно нормальное распределение в тексте. Этими оценками являются средняя арифметическая $\bar{x} = \theta_\alpha^*$ и медиана $\mathrm{Me} = \theta_\beta^*$. В. И. Романовский [30, кн. 1, с. 203] показал, что для средней арифметической дисперсия составляет

$$D(\theta_\alpha^*) = \sigma^2(\bar{x}) = \sigma^2/N,$$

а для медианы она равна

$$D(\theta_\beta^*) = \sigma^2(\mathrm{Me}) = \pi\sigma^2/(2N) = 1{,}57\sigma^2/N.$$

Так как концентрация значений θ_α^* вокруг $M(\theta^*)$ в полтора с лишним раза выше, чем концентрация значений θ_β^*, то средняя арифметическая является более эффективной оценкой математического ожидания, чем медиана.

§ 2. Оценка математического ожидания с помощью доверительного интервала и статистическая параметризация стилей

1. Доверительный интервал. Отвечая требованиям состоятельности, несмещенности и эффективности, точечная оценка θ^* имеет один серьезный недостаток: она не дает сведений о точности и надежности приближения к параметру θ генеральной лингвистической совокупности. Этот недостаток особенно ощутим тогда, когда число наблюдений мало́. Чтобы устранить этот недостаток, вводится иной вид оценки — *доверительный интервал*.

Для определения доверительного интервала необходимо найти такие случайные значения $\theta_н^*$, $\theta_в^*$, каждое из которых является функцией выборочных наблюдений $x_1, x_2, \ldots, x_N$, чтобы образуемый ими промежуток $(\theta_н^*, \theta_в^*)$ с вероятностью не менее $\mathfrak{p}$ покрывал неравенство $\theta_н^* < \theta < \theta_в^*$, в котором $\theta_н^*$ выступает в роли нижней, а $\theta_в^*$ — верхней границы доверительного интервала.

При оценке параметра θ с помощью доверительного интервала учитываются две величины.

Во-первых, необходимо указать ширину доверительного интервала $\theta_в^* - \theta_н^*$, половину которого составляет величина погрешности $\varepsilon = (\theta_в^* - \theta_н^*)/2$, характеризующая точность нашей оценки. Чем у́же интервал, т. е. чем меньше разность $\theta_в^* - \theta_н^*$, и следовательно, чем меньше ε, тем точнее оценка параметра θ.

Во-вторых, следует учитывать с какой вероятностью интервал $(\theta_н^*, \theta_в^*)$ покрывает параметр θ. Ведь нижняя $(\theta_н^*)$ и верхняя $(\theta_в^*)$ границы интервала суть случайные величины, зависящие от опыта: при многократном повторении опыта величина и положение интервала $(\theta_н^*, \theta_в^*)$ относительно θ будут меняться. Однако доверительный интервал должен быть построен так, чтобы доверительная вероятность (*надежность*) $\mathfrak{p}$ захвата им параметра θ была бы достаточно велика, а вероятность $\mathfrak{q}$, представляющая собой дополнение до единицы и указывающая на то, что θ не будет покрыт интервалом $(\theta_н^*, \theta_в^*)$, была бы соответственно мала (мы будем впредь называть $\mathfrak{q}$ *уровнем значимости*). Чем ближе вероятность $\mathfrak{p}$ к единице, а вероятность $\mathfrak{q}$ к нулю, тем меньше риск ошибиться в оценке параметра θ.

2. Определение доверительного интервала для математического ожидания нормально распределенной случайной величины при известном σ. Пусть случайная величина X имеет нормальное распределение в интересующей нас генеральной лингвистической совокупности, причем нам известно значение σ, но неизвестен параметр $M(X) = \mu$.

Для оценки μ естественно использовать среднюю арифметическую $\bar{x}$, которая на основании центральной предельной теоремы (см. гл. 6,

§ 4, п. 4) распределена нормально с параметрами $M(\bar{x}) = \mu$, $\sigma(\bar{x}) = \sigma/\sqrt{N}$.

Для того чтобы определить доверительный интервал неизвестного нам параметра $\mu = \theta$ по оценке $\bar{x} = \theta^*$, потребуем выполнения соотношения

$$P(|\bar{x} - \mu| < \varepsilon) = \mathfrak{p}, \tag{8.3}$$

где ε — допустимая погрешность, выступающая в качестве величины, обратной степени точности, а $\mathfrak{p}$, как уже говорилось, надежность, которая должна быть близка к единице.

Так как по предположению случайная величина X распределена нормально, то, на основании соотношения (6.125) выражение (8.3) можно переписать в виде

$$P(|\bar{x} - \mu| < z_{\mathfrak{p}}\,\sigma/\sqrt{N}) = 2\Phi(z_{\mathfrak{p}}) = \mathfrak{p}, \tag{8.4}$$

или

$$P(\bar{x} - z_{\mathfrak{p}}\,\sigma/\sqrt{N} < \mu < \bar{x} + z_{\mathfrak{p}}\,\sigma/\sqrt{N}) = \mathfrak{p}, \tag{8.5}$$

где индекс $\mathfrak{p}$ при z означает, что значение z берется в соответствии с надежностью (доверительной вероятностью) $\mathfrak{p}$.

Нетрудно заметить, что

$$\bar{x} - z_{\mathfrak{p}}\,\sigma/\sqrt{N} = \mu_{\text{н}}^* \tag{8.6}$$

является нижней, а

$$\bar{x} + z_{\mathfrak{p}}\,\sigma/\sqrt{N} = \mu_{\text{в}}^* \tag{8.7}$$

верхней границей доверительного интервала $(\mu_{\text{н}}^*, \mu_{\text{в}}^*)$. Этот интервал с надежностью $\mathfrak{p}$ и погрешностью $\varepsilon = z_{\mathfrak{p}}\sigma/\sqrt{N}$ покрывает математическое ожидание $\mu = M(X)$ нормально распределенной случайной лингвистической величины X в генеральной совокупности.

Расчет доверительного интервала производится следующим образом. Задавая определенную доверительную вероятность $\mathfrak{p}$, мы с помощью табл. III (см. стр. 365) определяем значение $z_{\mathfrak{p}}$, т. е. $z_{\mathfrak{p}} = x$. Подставив это значение в равенства (8.6) и (8.7), легко вычисляем значения верхней и нижней границ доверительного интервала.

Например, задав согласно правилу «трех сигм» надежность $\mathfrak{p} = 0{,}9973$ (т. е. $\mathfrak{p}/2 = 0{,}4986$), для нормально распределенной лингвистической величины получаем

$$\mu_{\text{н}}^* = \bar{x} - 3\sigma/\sqrt{N} \quad \text{и} \quad \mu_{\text{в}}^* = \bar{x} + 3\sigma/\sqrt{N}.$$

Погрешность в этом случае равна

$$\varepsilon(\mathfrak{p} = 0{,}997) = 3\sigma/\sqrt{N}.$$

Задавая по правилу «двух сигм» надежность $\mathfrak{p} = 0{,}954$, приходим к доверительным границам

$$\mu_{\text{н}}^* = \bar{x} - 2\sigma/\sqrt{N}, \quad \mu_{\text{в}}^* = \bar{x} + 2\sigma/\sqrt{N}$$

при величине погрешности, равной

$$\varepsilon(\mathfrak{p} = 0{,}954) = 2\sigma/\sqrt{N}.$$

Наконец, удовлетворяясь надежностью $\mathfrak{p} = 0{,}683$, мы получаем еще более узкий доверительный интервал с границами

$$\mu_{\text{н}}^* = \bar{x} - \sigma/\sqrt{N}, \quad \mu_{\text{в}}^* = \bar{x} + \sigma/\sqrt{N},$$

причем погрешность равна

$$\varepsilon(\mathfrak{p} = 0{,}683) = \sigma/\sqrt{N} = \sigma(\bar{x}).$$

Из всего сказанного становится ясным, что *при одном и том же объеме выборки N с увеличением надежности $\mathfrak{p}$ уменьшается точность* (т. е. значение погрешности ε становится бо́льшим) и, наоборот, *с уменьшением $\mathfrak{p}$ увеличивается точность* (т. е. численное значение ε становится меньшим). Если сохранять постоянным значение надежности $\mathfrak{p}$ и увеличивать объемы выборки N, то можно увеличить точность, и наоборот, при постоянном $\mathfrak{p}$ с уменьшением N точность уменьшить. Наконец, одновременного увеличения надежности и точности можно достичь только путем увеличения объема выборки N. Все эти соображения будут иметь принципиальное значение при определении необходимых объемов лингвистических выборок.

3. **Определение доверительного интервала для $M(X)$ с помощью распределения Стьюдента.** В лингвистической практике редко встречаются ситуации, при которых известно численное значение среднего квадратического отклонения в генеральном распределении. Чаще всего и математическое ожидание, и дисперсия, и среднее квадратическое отклонение остаются неизвестными. Поэтому необходимы такие процедуры поиска интервальных оценок параметров теоретического распределения, которые опирались бы только на значения средней $\bar{x}$, опытной дисперсии s^2 и стандарта s (соответственно $\hat{s}^2$, $\hat{s}$), получаемых из частных выборок, взятых из генеральной лингвистической совокупности.

При этом интервальная оценка математического ожидания $M(X) = \mu$ достигается путем применения распределения Стьюдента.

В гл. 6 (см. § 3, п. 4 и § 4, п. 4) было показано, что в тех случаях, когда величина X распределена нормально с математическим ожиданием $M(X) = \mu$ и средним квадратическим отклонением $\sigma(\bar{x}) = \sigma/\sqrt{N}$, величина

$$Z = \frac{\bar{x} - \mu}{\sigma}\sqrt{N} \tag{8.8}$$

дает нормированное нормальное распределение со средней, равной нулю, и дисперсией, равной единице.

Однако при определении доверительного интервала распределением Z воспользоваться нельзя, поскольку величина σ неизвестна. Заменим σ стандартом

$$\hat{s} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}}$$

и перепишем (8.8) в виде

$$t = \frac{\bar{x} - \mu}{\hat{s}} \sqrt{N}. \qquad (8.9)$$

В работах по математической статистике [63, с.198] доказывается, что плотность вероятности распределения значений t задается выражением

$$f_\nu(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \qquad (8.10)$$

где $\nu = N - 1$, $-\infty < t < \infty$, а Γ — так называемая гамма-функции [9, с. 161 и сл.]. Из (8.10) видно, что распределение величины t не зависит от неизвестных параметров $M(X) = \mu$ и $D(X) = \sigma^2$, а зависит лишь от величины ν, называемой *числом степеней свободы* (см. ниже § 3, п.1), которая представляет собой численность $N - 1$ независимых значений случайной величины X в выборке из N испытаний.

Интегрируя (8.10) в пределах от $-\infty$ до t_p, можно найти вероятность $P(t < t_p)$ случайных значений t, меньших, чем заданное значение t_p:

$$P(|t| < t_p) = \int_{-\infty}^{t_p} f_\nu(x)\, dx. \qquad (8.11)$$

Описанный закон распределения носит название **закона распределения Стьюдента** (t-распределения) с ν степенями свободы (в литературе он иногда называется распределением малых выборок).

Так как при больших значениях N выборочная дисперсия s^2 и стандарт s мало чем отличаются от теоретических параметров σ^2 и σ, то при больших $\nu = N - 1$ величина t, приближаясь к Z, получает нормальное распределение. Однако когда $\nu = N - 1$ мало, то t, сильно отличаясь от Z, не подчиняется нормальному распределению. Эти особенности t-распределения показаны на рис. 57. Сохраняя колоколообразную и симметричную относительно начала координат форму, кривая распределения Стьюдента при малых значениях N и ν гораздо медленнее сближается с осью абсцисс, чем кривая нормального распределения. Поэтому вероятность значений t, попадающих в критическую область или, иными словами, превышающих по абсолютной величине заданный предел t_p, гораздо больше, чем вероятность значений Z, превышающих установленный предел $z_p = t_p$ (рис. 57). Однако при $N \to \infty$ кривая $f(t)$ совпадает с кривой нормального распределения.

Учитывая равенство (8.11) и симметрию кривой $f_\nu(t)$, можно легко прийти к вероятности того, что величина t будет находиться в заданных пределах $(-t_p, t_p)$. Эта доверительная вероятность (надежность) равна

$$p = P(-t_p < t < t_p) = P(|t| < t_p). \qquad (8.12)$$

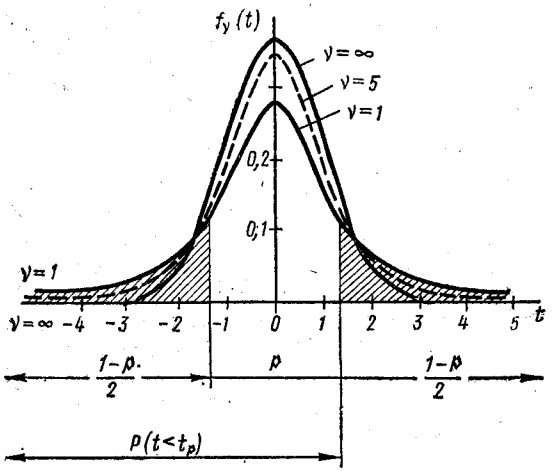


Рис. 57

Аналогичным образом уровень значимости составляет

$$q = P(|t| \geqslant t_p). \qquad (8.13)$$

При решении лингвистических задач с помощью распределения Стьюдента доверительные вероятности p определяются по табл. IV (см. стр. 366, 367), строки которой дают заданные значения t, а столбцы—заданные величины $\nu = N - 1$. На пересечении строк и столбцов находятся соответствующие значения доверительной вероятности.

Распределение Стьюдента легко может быть использовано для интервальной оценки математического ожидания $M(X) = \mu$ лингвистической случайной величины, относительно которой наперед известно, что она распределена нормально, но параметры которой $D(X) = \sigma^2$ и σ остаются неизвестными.

Действительно, подставляя значение t в неравенство $-t_p < t < t_p$, вероятность которого задана наперед, имеем

$$-t_p < \frac{\bar{x} - \mu}{\hat{s}/\sqrt{N}} < t_p, \qquad (8.14)$$

или

$$\bar{x}-t_{\mathfrak{p}}\,\hat{s}/\sqrt{N}<\mu<\bar{x}+t_{\mathfrak{p}}\,\hat{s}/\sqrt{N}. \qquad (8.15)$$

Это неравенство, равносильное (8.12), имеет вероятность

$$P\left(\bar{x}-t_{\mathfrak{p}}\,\hat{s}/\sqrt{N}<\mu<\bar{x}+t_{\mathfrak{p}}\,\hat{s}/\sqrt{N}\right)=\mathfrak{p}=2\int_{0}^{t_{\mathfrak{p}}} f_{\nu}(x)\,dx. \qquad (8.16)$$

Нетрудно заметить, что вероятность $\mathfrak{p}$, легко получаемая из таблицы значений $P\,(|t|\geqslant t_{\mathfrak{p}})$ (см. табл. IV), является надежностью нашего утверждения о том, что t не выйдет из доверительного интервала с нижней границей

$$\mu_{\text{н}}^{*}=\bar{x}-t_{\mathfrak{p}}\,\hat{s}/\sqrt{N} \qquad (8.17)$$

и верхней границей

$$\mu_{\text{в}}^{*}=\bar{x}+t_{\mathfrak{p}}\,\hat{s}/\sqrt{N}, \qquad (8.18)$$

где член $t_{\mathfrak{p}}\,\hat{s}/\sqrt{N}=t_{\mathfrak{p}}\hat{s}\,(\bar{x})$ оценивает погрешность ε.

Только что полученная интервальная оценка очень похожа на оценку, приведенную в п. 2. Различие состоит в том, что вместо теоретического среднего квадратического отклонения σ здесь используется выборочная величина $\hat{s}$, а вместо $z_{\mathfrak{p}}$ применяются $t_{\mathfrak{p}}$. В тех случаях, когда выборка $N=\nu+1$ велика (например, когда $N>30$), $\hat{s}\approx\sigma$, а $t_{\mathfrak{p}}\approx z_{\mathfrak{p}}$, и поэтому доверительный интервал, полученный с помощью t-распределения, близок к интервалу, вычисленному в п. 2. В тех случаях, когда $N=\nu+1$, мало, t-интервал заметно шире z-интервала. Разумеется, это не является недостатком самого распределения Стьюдента. Причина кроется в малом объеме выборки: чем меньше объем выборки, тем меньше информации о генеральной совокупности, в том числе и о параметре $M(X)=\mu$, содержит выборочное распределение Стьюдента. Поэтому и ширина доверительного интервала должна быть здесь больше по сравнению с тем интервалом, который получен на основании сведений о распределении самой генеральной совокупности.

4. Математическое ожидание как статистический параметр стиля. Одним из важных вопросов квантитативной лингвистики является выявление объективных статистических признаков для отдельных разновидностей языка (стилей, подъязыков, жанров, авторского стиля). Эта проблема исследовалась коллективом языковедов, руководимым В. И. Перебейнос [34]. В частности, была сделана попытка разграничить жанры и стили современного украинского языка с точки зрения частоты употребления в них глагольных словоформ.

Это исследование строилось следующим образом. Из современных художественных, общественно-политических и научно-технических текстов было извлечено 280 выборок по 500 словоупотребле-

Таблица 8.1

Статистические характеристики употребления глаголов в различных стилях современного украинского языка

Жанры и стили украинского литературного языка	Количество выборок N	$\bar{x}=\bar{F}$	$\hat{s}$	$\hat{s}(\bar{x})$	$\mathfrak{p}=0,95$	$t_{\mathfrak{p}}=2,00$	$\mathfrak{p}=0,95$	$z_{\mathfrak{p}}=1,96$	$\mathfrak{p}=0,996$	$t_{\mathfrak{p}}=3,00$
					$\varepsilon=t_{\mathfrak{p}}\hat{s}(\bar{x})$	$(\mu_{\text{н}}^{*}, \mu_{\text{в}}^{*})$	$\varepsilon=z_{\mathfrak{p}}\sigma(\bar{x})$	$(\mu_{\text{н}}^{*}, \mu_{\text{в}}^{*})$	$\varepsilon=3\sigma(\bar{x})$	$(\mu_{\text{н}}^{*}, \mu_{\text{в}}^{*})$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Язык художественной литературы:										
драма	60	90,5	15,36	1,98	3,96	86,54—94,46	3,88	86,62—94,38	5,94	84,56—96,44
проза	60	91,2	13,45	1,73	3,46	87,74—94,66	3,39	87,81—94,59	5,19	86,01—96,39
поэзия	50	82,1	10,44	1,48	2,97	79,13—85,07	2,90	79,20—85,00	4,44	77,66—86,54
Общественно-политическая проза	60	48,0	10,55	1,35	2,70	45,3—50,7	2,65	45,35—50,65	4,05	43,95—52,05
Научно-техническая проза	50	61,7	9,92	1,40	2,80	58,9—64,5	2,74	58,96—64,44	4,20	57,50—65,90

ний в каждой (количество выборок по стилям и жанрам показано в столбце (2) табл. 8.1]. Для каждого жанра и стиля была вычислена средняя частота $\bar{F} = \bar{x}$ глагольных словоупотреблений, а также стандарты $s \approx \hat{s}$ и $s(\bar{x}) = s/\sqrt{N} \approx \hat{s}(\bar{x})$.

Поскольку дисперсия и среднее квадратическое генеральных текстовых совокупностей по украинским стилям остаются неизвестными, вычисление доверительного интервала, в котором находятся величины $M(F)$, следует осуществлять с помощью распределения Стьюдента. При этом, опираясь на данные предшественников [32а, с. 43—45, 104—107], можно предположить, что частоты глагольных словоупотреблений распределены нормально.

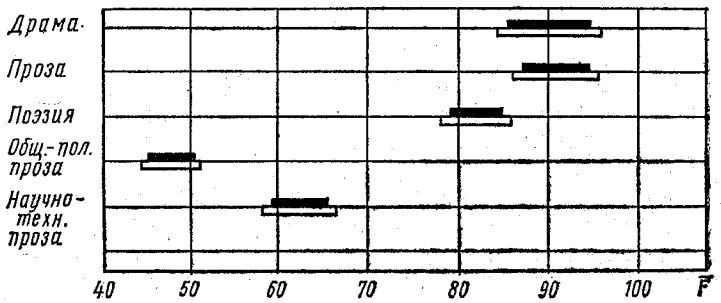


Рис. 58. Доверительные интервалы математического ожидания частоты глагольных словоупотреблений в украинской драме, прозе и поэзии:
▬ — доверительный интервал при $\mathfrak{p} = 0{,}95$;
▭ — доверительный интервал при $\mathfrak{p} > 0{,}996$ (правило «трех сигм»)

Если удовлетвориться надежностью наших утверждений, равной 95% ($\mathfrak{p} = 0{,}95$), то, согласно данным табл. 8.1, при $\nu = 60 - 1 = 59$ имеем $t_{\mathfrak{p}} = 2{,}00$, а при $\nu = 50 - 1 = 49$ получаем $t_{\mathfrak{p}} = 2{,}01$.

В этом случае величина погрешности при определении $M(\bar{F})$ глагольных словоупотреблений в выборках из украинской драмы составляет $\varepsilon = 2{,}00 \cdot 1{,}98 = 3{,}96$, а границы доверительного интервала, в силу равенств (8.17) и (8.18),

$$\mu_{\text{н}}^{*} = 90{,}5 - 3{,}96 = 86{,}54,\ \mu_{\text{в}}^{*} = 90{,}5 + 3{,}96 = 94{,}46.$$

Аналогичные значения погрешности и границ доверительных интервалов относительно других жанров показаны в столбцах (6) и (7).

Если положить, что среднее квадратическое отклонение σ равно полученному экспериментальным путем стандарту s, то интервал, покрывающий с надежностью $\mathfrak{p} = 0{,}95$ величину $M(\bar{F})$, несколько сузится [см. столбец (9)].

Как показывают данные табл. 8.1 и рис. 58, доверительные интервалы $M(\bar{F})$ глагольных словоупотреблений в украинской драме и прозе частично накладываются друг на друга. Что касается других жанров и стилей, то там этих наложений нет. Отсюда можно сделать вывод, что частота глагольных словоупотреблений является тем статистическим параметром, который обособляет украинские драматические и прозаические тексты от поэтических, общественно-политических и научно-технических текстов, а также различает между собой последние два стиля (о том, насколько точны эти заключения, сделанные при условии точечной оценки среднего квадратического отклонения, будет сказано в следующем параграфе).

Лингвистические выводы, к которым мы только что пришли, имеют сравнительно малую надежность — всего 95%. Если же необходимо увеличить надежность этих выводов, приблизив их к 100%-ному утверждению, то следует произвести расчет границ доверительного интервала, опираясь на правило «трех сигм». Платой за это увеличение надежности будет заметный рост погрешности ε [см. столбец (10) табл. 8.1] и расширение доверительного интервала [столбец (11)]. Это расширение приводит к тому, что правый конец доверительного интервала поэзии накладывается на левый конец интервалов драмы и прозы. Поэтому вывод о том, что частота глагольных словоупотреблений является статистическим параметром, отграничивающим украинскую драму и прозу от поэзии, оказывается необоснованным. Частоту глаголов следует теперь рассматривать в качестве параметра, различающего только научно-техническую, общественно-политическую и художественную украинскую речь.

Пользуясь только что описанной методикой, В. И. Перебейнос выделила 74 языковых признака (фонемы, аффиксы, морфологические и синтаксические классы, длина слова и предложения и т. п.), частота которых выступает в роли статистического параметра стилей, жанров и авторской манеры в украинском языке. Выясняется, что количество признаков каждого уровня языковой структуры, способных разграничивать языковые разновидности, зависит не от уровня, к которому относится данный признак, а от сопоставляемых разновидностей. Чем больше общих черт имеют сопоставляемые разновидности, тем меньше признаков их различает. Например, на буквенно-фонемном уровне наименьшее количество характеристик различает общественно-политическую и научно-техническую речь. Для каждой пары разновидностей существует свой уровень, лингвистические признаки которого лучше всего разграничивают сравниваемые стили и жанры. Более родственные разновидности (например драма, поэзия, проза) обнаруживают заметные расхождения на синтаксическом уровне, а менее родстенные, кроме того, и на лексическом (ср. противопоставление беллетристики, с одной стороны, и общественно-политической и научно-технической прозы — с другой).

Как показывает работа [34], статистическое моделирование стилей не только хорошо согласуется с их интуитивными описаниями, но помогает выявить недоступные для прямого лингвистического наблюдения факты.

§ 3. Доверительный интервал для дисперсии и среднего квадратического отклонения

В предыдущем параграфе мы пользовались точечной оценкой среднего квадратического отклонения. Однако, как уже говорилось, такая оценка не дает сведений о том, насколько близок стандарт s к самому значению σ.

В тех случаях, когда в генеральной лингвистической совокупности имеет место нормальное распределение величины X и известна выборочная дисперсия s, этого недостатка можно избежать, используя для оценок $D(X) = \sigma^2$ и σ доверительный интервал, который вычисляется с помощью распределения χ^2 Пирсона.

1. Распределение χ^2 Пирсона. Пусть имеется N независимых нормированных случайных величин

$$Z_1 = \frac{x_1 - \mu}{\sigma}, \quad Z_2 = \frac{x_2 - \mu}{\sigma}, \ldots, Z_N = \frac{x_N - \mu}{\sigma}, \tag{8.19}$$

каждая из которых распределена нормально и обладает параметрами $M(Z) = 0$ и $\sigma(Z) = 1$. Сумму квадратов этих величин обозначим через χ^2:

$$\chi^2 = \sum_{i=1}^{N} Z_i^2 \quad (0 \leqslant \chi^2 < \infty).$$

Тогда плотность вероятности χ^2 определяется выражением

$$P_\nu(\chi^2) = \frac{1}{2^{\nu/2}\,\Gamma(\nu/2)} \cdot (\chi^2)^{\nu/2-1} \cdot e^{-\chi^2/2},$$

где $\nu = N$ — число независимых слагаемых в последовательности (8.19), а $\Gamma(\nu/2)$ — гамма-функция; см. § 2, п. 3, а также [36, с. 200 и сл.].

Интегральная функция распределения Пирсона имеет вид

$$P(\chi^2 > \chi_0^2) = \int_{\chi_0^2}^{\infty} P_\nu(\chi^2)\, d\chi^2.$$

При $\nu \to \infty$ распределение величины χ^2 асимптотически приближается к нормальному распределению с параметрами $M(\chi^2) = \nu$, $D(\chi^2) = 2\nu$, $\sigma = \sqrt{2\nu}$ [61, с. 119].

Правильность нахождения $P(\chi^2 > \chi_0^2)$ зависит от корректного определения величины ν, называемой *числом степеней свободы*. Число степеней свободы определяет то количество сведений, которое остается свободным после использования всей совокупности сведений для определения некоторой статистической характеристики. Поясним это важное для лингво-статистики понятие на следующем примере.

Будем рассматривать русские прозаические тексты в качестве генеральной совокупности, из которой последовательно выбираются отрывки по 10 тыс. словоупотреблений. В каждом отрывке отмечается количество употреблений слова *море* (частота употреблений этого слова является случайной величиной X, принимающей значения $x_1, x_2, \ldots, x_N$). Пусть в результате этих наблюдений получены следующие сведения:

$$x_1 = 2,\ x_2 = 3,\ x_3 = 0,\ x_4 = 4,\ x_5 = 5,\ x_6 = 3,\ \ldots\ .$$

Число отрывков, в которых производится наблюдение, у нас не фиксировано (оно может быть и бесконечным). Каждая случайная величина может принимать любое значение. Таким образом, число степеней свободы теоретически равно здесь количеству взятых нами значений случайной величины X, т. е. $\nu = N$ (при этом мы не должны забывать о том, что нами всегда могут быть взяты и другие неучтенные значения X).

Теперь зафиксируем число интересующих нас отрывков величиной $N = 6$ и определим среднюю арифметическую частоты слова *море* относительно этих отрывков:

$$\bar{x} = \sum_{i=1}^{N} x_i/N = (2 + 3 + 0 + 4 + 5 + 3)/6 = 2{,}83.$$

Фиксирование числа выборок и суммирование частот появления в них слова *море* накладывает на нашу совокупность сведений, содержащуюся в таблице, одну линейную связь. Действительно, зная сумму частот $\sum_{i=1}^{N} x_i$ слова *море*, а также частоты этого слова в каждом из $N - 1$ отрывков, мы всегда можем вычислить частоту контрольного слова в N-й выборке:

$$x_N = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N-1} x_i.$$

Таким образом, значения частот слова *море* могут варьироваться в каждом из $N - 1$ отрывков, но эти изменения каждый раз будут предопределять частоту в N-й выборке, если общая сумма частот уже известна. Иными словами, после определения величин $N = \sum_{i=1}^{N} x_i$ мы наложили на нашу статистическую совокупность $x_1, x_2, \ldots, x_N$ о д н у связь. Количество свободных сведений (число степеней свободы) уменьшилось при этом на единицу: $\nu = N - 1$.

Эта связь учитывается и при определении других выборочных характеристик — например при вычислении выборочной дисперсии и стандарта s, при расчете которых сумма квадратов отклонений делится не на N, а на $N - 1$ (см. § 1, п. 2). Как уже неоднократно указывалось, при больших N ($N > 30$) указанной поправкой можно пренебречь.

В биномиальном распределении и в распределении Пуассона должны быть учтены д в е связи: во-первых, связь, возникающая при суммировании наблюдаемых частот, о которой мы только что говорили, а во-вторых, связь, образующаяся при определении теоретических параметров (p — для биномиального распределения и λ — для распределения Пуассона). Таким образом, число степеней свободы здесь равно $\nu = N - 2$.

В случае нормального распределения приходится учитывать уже т р и связи: связь, образующуюся при суммировании наблюдаемых частот, и связи, содержащиеся в величинах $\bar{x}$ и s, соответственно оценивающих $M(X)$ и σ. Таким образом, в этом случае $\nu = N - 3$.

Однако если при сопоставлении нормального распределения с выборочным величины $\bar{x}$ и s оказываются уже известными до опыта, то тогда должна быть учтена всего лишь одна связь, образующаяся при суммировании наблюдаемых частот. В связи с этим здесь число степеней свободы $\nu = N - 1$.

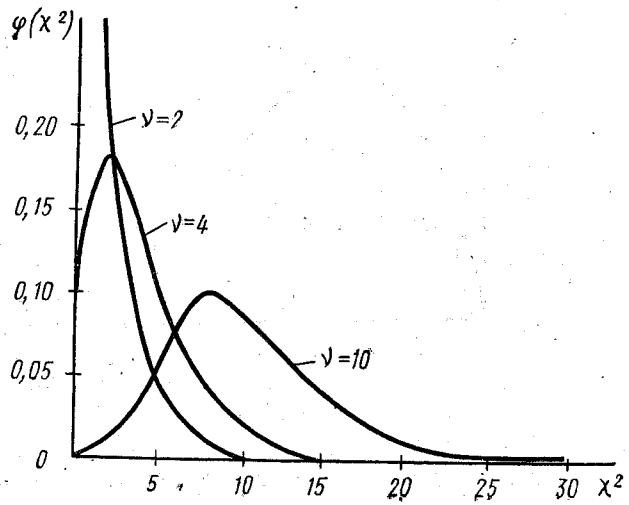


Рис. 59

Итак, число степеней свободы определяется из разности

$$\nu = N - l, \qquad (8.20)$$

где N — число наблюдений (сведений), а l — число линейных связей, налагающихся на эти наблюдения при данной статистической процедуре.

От числа степеней свободы зависит вид кривых χ^2 распределения. Как видно из рис. 59, эти кривые асимметричны, причем степень асимметрии уменьшается с увеличением числа степеней свободы. При $\nu \to \infty$ график χ^2 совпадает с кривой нормального распределения.

При определении доверительных интервалов для величин $D(X) = \sigma^2$ и σ используется выборочная дисперсия $\hat{s}^2$, связанная с величинами σ^2 и χ^2 следующей зависимостью:

$$\hat{s}^2 = \sigma^2 \chi^2 / \nu, \qquad (8.21)$$

причем $\nu = N - 1$.

Распределение $\hat{s}^2$ подробно описано в работе [61, с. 139 — 144]. Отметим лишь, что оно имеет параметры

$$M(\hat{s}^2) = \sigma^2, \quad D(\hat{s}^2) = 2\sigma^4/\nu.$$

Поскольку при $\nu \to \infty$ распределение χ^2 асимптотически приближается к нормальному, при этих условиях распределение $\hat{s}^2$ также должно приближаться к нормальному распределению.

Переходя к определению доверительного интервала для σ^2 и σ, перепишем равенство (8.21) в виде

$$\chi^2 = \nu\hat{s}^2/\sigma^2. \qquad (8.22)$$

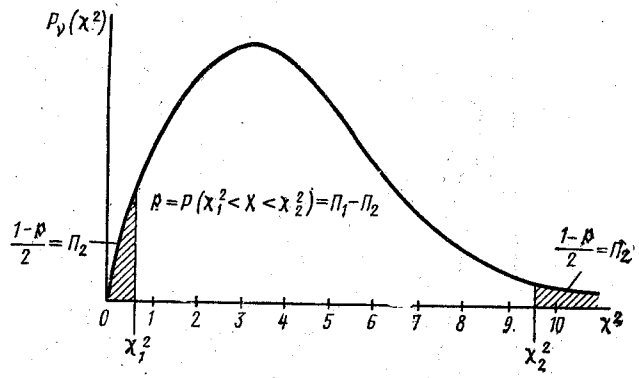


Рис. 60

Теперь, задав надежность $\mathfrak{p}$, найдем в таблице распределения χ^2 такие две численные границы χ_1^2 и χ_2^2, для которых выполнялось бы соотношение

$$P(\chi_1^2 < \nu\hat{s}^2/\sigma^2 < \chi_2^2) = \mathfrak{p}. \qquad (8.23)$$

Границы χ_1^2 и χ_2^2 можно выбрать бесконечным числом способов: они могут быть сдвинуты по оси абсцисс (рис. 60) влево или вправо на любую величину, лишь бы число $\chi^2 = \nu\hat{s}^2/\sigma^2$ находилось внутри интервала (χ_1^2, χ_2^2) и значение $\mathfrak{p}$ оставалось бы неизменным.

Чтобы фиксировать положение границ χ_1^2 и χ_2^2, вводят односторонний критерий значимости, согласно которому

$$P(\chi^2 \leqslant \chi_1^2) = P(\chi^2 \geqslant \chi_2^2) = \frac{1-\mathfrak{p}}{2}.$$

При этом условии надежность наших рассуждений остается равной $\mathfrak{p}$. Действительно,

$$P(\chi_1^2 < \nu s^2/\sigma^2 < \chi_2^2) = 1 - P(\chi^2 \leqslant \chi_1^2) - P(\chi^2 \geqslant \chi_2^2) =$$

$$= 1 - \left(\frac{1-\mathfrak{p}}{2} + \frac{1-\mathfrak{p}}{2}\right) = \mathfrak{p}. \qquad (8.24)$$

Преобразуя выражение, стоящее в левой части равенства (8.24), таким образом, что

$$P\,(\nu\hat{s}^2/\chi_2^2 < \sigma^2 < \nu\hat{s}^2/\chi_1^2) = \mathfrak{p},$$

мы получаем с надежностью $\mathfrak{p}$ доверительный интервал для дисперсии σ^2, имеющий нижнюю границу

$$\sigma_{\text{н}}^{2*} = \nu\hat{s}^2/\chi_2^2 \qquad (8.25)$$

и верхнюю границу

$$\sigma_{\text{в}}^{2*} = \nu\hat{s}^2/\chi_1^2. \qquad (8.26)$$

Границы доверительного интервала для среднего квадратического отклонения соответственно равны

$$\sigma_{\text{н}}^{*} = \sqrt{\nu}\,\hat{s}/\chi_2 \qquad (8.27)$$

и

$$\sigma_{\text{в}}^{*} = \sqrt{\nu}\,\hat{s}/\chi_1. \qquad (8.28)$$

Для нахождения значений χ_1^2 и χ_2^2 следует воспользоваться табл. V на стр. 368. Столбцы этой таблицы указывают $(1 + \mathfrak{p})/2$ и $(1 - \mathfrak{p})/2$, а строки — число степеней свободы. Нужные значения односторонних границ χ_1^2 и χ_2^2 находятся на пересечении строк и столбцов. Подставив эти значения в формулы (8.25) — (8.28), получаем верхнюю и нижнюю оценки неизвестных параметров $D(X) = \sigma^2$ и $\sqrt{D\,(X)} = \sigma$.

2. Определение существенности расхождения частот глагола в украинской драме и поэзии с использованием интервальной оценки. Решая вопрос о существенности статистических расхождений между украинской драмой и прозой [см. столбец (9) табл. 8.1 на стр. 275], мы пользовались точечной оценкой среднего квадратического отклонения. Как уже говорилось, эта оценка не дает сведений о том, как далеко отстоит от наблюдаемого стандарта $\hat{s}$ или s среднее квадратическое σ. Поэтому в тех случаях, когда сопоставляемые лингвистические признаки дают близкие средние частоты $\bar{F} = \bar{x}$, как это имеет место при сравнении употребительности глаголов в украинской драме и прозе, использование точечных оценок σ может привести к ошибочным лингвистическим выводам. В связи с этим снова обратимся к вычислению доверительного интервала для средних частот глагола в украинских жанрах украинской прозы, пользуясь при этом не точечными, а интервальными оценками среднего квадратического $\sigma\,(\bar{x})$.

По условию задачи, имеем для драматических текстов $\nu = 60 - 1 = 59$, $\bar{F} = \bar{x} = 90{,}5$, $\hat{s}(\bar{x}) = 1{,}98$, а для поэтических $\nu = 50 - 1 = 49$, $\bar{F} = \bar{x} = 82{,}1$, $\hat{s}(\bar{x}) = 1{,}48$. И в том, и в другом случае $\mathfrak{p} = 0{,}95$, откуда $(1 - \mathfrak{p})/2 = 0{,}025$, $(1 + \mathfrak{p})/2 = 0{,}975$. Учитывая, что для драмы $\nu = 59$, из табл. V находим

$$\chi_1^2 = \chi^2_{(1+\mathfrak{p})/2;\,\nu} = \chi^2_{0{,}975;\,59} = 39{,}66,$$

$$\chi_2^2 = \chi^2_{(1-\mathfrak{p})/2;\,\nu} = \chi^2_{0{,}025;\,59} = 82{,}18.$$

Соответственно для поэзии, где $\nu = 49$, имеем

$$\chi_1^2 = \chi^2_{0{,}975;\,49} = 31{,}56, \quad \chi_2^2 = \chi^2_{0{,}025;\,49} = 70{,}22.$$

Таким образом, нижняя и верхняя границы доверительного интервала для σ в драматических текстах составляют

$$\sigma_{\text{н}}^{*} = 1{,}98\sqrt{59/82{,}18} = 1{,}68; \quad \sigma_{\text{в}}^{*} = 1{,}98\sqrt{59/39{,}66} = 2{,}43,$$

а для поэзии эти границы равны

$$\sigma_{\text{н}}^{*} = 1{,}48\sqrt{49/70{,}22} = 1{,}24; \quad \sigma_{\text{в}}^{*} = 1{,}48\sqrt{49/31{,}56} = 1{,}84.$$

Из табл. 8.1 и рис. 58 видно, что хотя при точечной оценке σ доверительный интервал употребительности глагола в поэзии не накладывается на доверительный интервал глагола в драме, верхняя граница первого интервала очень близко подходит к нижней границе второго.

Теперь посмотрим, сохраняется ли этот разрыв между доверительными интервалами при использовании интервальной оценки σ. Для этого определим нижнюю границу $\mu_{\text{н}}^{*}$ доверительного интервала $\bar{F} = \bar{x}$ глагола в драме и верхнюю границу аналогичного интервала в поэзии. Само собой разумеется, что и в том, и в другом случае следует пользоваться верхней границей интервальной оценки среднего квадратического $\sigma_{\text{в}}^{*}$. Тогда для драматических текстов получаем

$$\mu_{\text{н}}^{*} = \bar{x} - z_{\mathfrak{p}}\sigma_{\text{в}}^{*} = 90{,}5 - 1{,}96 \cdot 2{,}43 = 85{,}73,$$

а для поэзии соответственно имеем

$$\mu_{\text{в}}^{*} = 82{,}1 + 1{,}96 \cdot 1{,}86 = 85{,}71.$$

Полученные результаты показывают, что нижняя часть доверительного интервала $\bar{F} = \bar{x}$ глаголов в украинской драме по существу соприкасается с верхней частью соответствующего доверительного интервала в поэзии. Поэтому благоразумнее воздержаться от сделанного с надежностью в 95% вывода § 2, п. 4 о том, что частота глаголов может рассматриваться в качестве статистико-стилистического параметра, надежно отграничивающего язык украинской драмы от языка поэзии.

§ 4. Доверительные интервалы для вероятности качественного лингвистического признака

В лингвистической практике постоянно приходится применять интервальную оценку вероятности отдельных единиц — фонем, графем, слогов, морфем, словоформ и т. д. При осуществлении этой оценки интересующая нас лингвистическая единица, например глагольная словоформа, рассматривается в виде качественного альтернативного признака A. Все остальные лингвистические единицы (в нашем случае — все неглагольные словоформы) квалифицируются здесь как качественный признак $\bar{A}$, т. е. не A.

В гл. 6 (см. § 1, п. 3) было показано, что вероятность p альтернативной величины A имеет биномиальное распределение. Однако определение интервальной оценки p при биномиальном распределении этой вероятности связано с громоздкими расчетами, опирающимися на довольно сложный математический аппарат. Чтобы обойти эти затруднения, пользуются более простыми приемами определения интервальной оценки p.

1. Интервальная оценка вероятности p с помощью нормального распределения. Пусть вероятность p альтернативного лингвистического признака A не слишком близка к нулю и к единице, а число наблюдений N достаточно велико (такая ситуация имеет обычно место при статистических исследованиях в области грамматики, фонетики и фонологии). Вероятность p неизвестна, и ее нужно оценить через получаемую в опыте частость $f = F/N$.

В гл. 6 (см. § 4) было показано, что распределение величины f близко к нормальному Теоретически можно предполагать, что параметрами этого распределения служат величины

$$M(F/N) = M(f) = p \quad \text{и} \quad \sigma(f) = \sqrt{p(1-p)/N}$$

(см. гл. 6, § 3, п. 4) При бесконечном увеличении числа испытаний (т. е. объема выборки) предельным распределением нормированной частости

$$\frac{M(f)-f}{\sigma} = \frac{p-f}{\sqrt{p(1-p)/N}}$$

является нормальное распределение

$$\left(\left|\frac{p-f}{\sigma}\right| < z_{\mathfrak{p}}\right) = 2\Phi(z) = \mathfrak{p}, \qquad (8.29)$$

где $z_{\mathfrak{p}}$ — величина, значение которой соответствует заданной надежности $\mathfrak{p}$ при $\nu = \infty$ (см. табл. VI на стр. 369).

От неравенства

$$\left|\frac{p-f}{\sigma}\right| < z_{\mathfrak{p}} \qquad (8.30)$$

перейдем к двойному неравенству

$$-z_{\mathfrak{p}} < (p-f)/\sigma < z_{\mathfrak{p}}, \text{ или } f - \sigma z_{\mathfrak{p}} < p < f + \sigma z_{\mathfrak{p}},$$

где

$$\sigma = \sqrt{p(1-p)/N}. \qquad (8.31)$$

Заменив в (8.31) вероятность p на полученное из опыта значение частости f и следуя рассуждениям п. 2, § 2, приходим к доверительному интервалу вероятности, который имеет вид

$$f - z_{\mathfrak{p}}\sqrt{f(1-f)/N} < p < f + z_{\mathfrak{p}}\sqrt{f(1-f)/N}. \qquad (8.32)$$

При этом нижняя оценка параметра p равна

$$p_{\text{н}}^* = f - z_{\mathfrak{p}}\sqrt{f(1-f)/N} = f - \varepsilon, \qquad (8.33)$$

а верхняя составляет

$$p_{\text{в}}^* = f + z_{\mathfrak{p}}\sqrt{f(1-f)/N} = f + \varepsilon, \qquad (8.34)$$

где величина ε указывает на погрешность при определении доверительного интервала для p:

$$z_{\mathfrak{p}}\sqrt{f(1-f)/N} = z_{\mathfrak{p}}\sigma = \varepsilon. \qquad (8.35)$$

Применим только что описанную процедуру к конкретной лингвистической задаче.

В молдавском публицистическом тексте длиной в 200 тыс. словоупотреблений встретилось 31286 глагольных форм [7, с. 159]. Нужно с надежностью в 95% определить доверительные границы вероятности появления во взятом тексте глагольного словоупотребления.

Здесь $N = 200000$, $F = 31286$, $f = F/N = 0{,}1564$. По табл. VI находим, что $z_{\mathfrak{p}} = 1{,}96$ при $\mathfrak{p} = 0{,}95$ и при $\nu = \infty$; затем по формуле (8.35) определим погрешность

$$\varepsilon = 1{,}96\sqrt{0.1564 \cdot 0{,}8436/200\,000} \approx 0{,}0016.$$

Подставляя все эти данные в равенства (8.33) и (8.34), получаем значения

$$p_{\text{н}}^* \quad 0.1564 - 0{,}0016 = 0{,}1548, \quad p_{\text{в}}^* = 0{,}1564 + 0{,}0016 = 0{,}1580$$

нижней и верхней границ доверительного интервала, в котором с надежностью в 95% находится истинная вероятность молдавских глагольных словоупотреблений.

Приближенная оценка вероятности p с помощью опытной частости f всегда связана с ошибкой, величина которой тем больше, чем меньше объем выборки. Более точную интервальную оценку можно получить, решая неравенство (8.30) относительно p. Для этого указанное неравенство запишем в виде

$$p - f < z_{\mathfrak{p}}\sqrt{p(1-p)/N}. \qquad (8.36)$$

Затем обе части неравенства возведем в квадрат и перенесем все его члены в левую часть:

$$p^2\left(1 + \frac{z_{\mathfrak{p}}^2}{N}\right) - p\left(2f + \frac{z_{\mathfrak{p}}^2}{N}\right) + f^2 < 0. \qquad (8.37)$$

Приравняв левую часть неравенства нулю и решая квадратное уравнение относительно p, получаем

$$p_1 = p_{\text{н}}^* = \frac{Nf + \frac{1}{2}z_{\mathfrak{p}}^2 - z_{\mathfrak{p}}\sqrt{Nf(1-f) + \frac{1}{4}z_{\mathfrak{p}}^2}}{N + z_{\mathfrak{p}}^2}, \qquad (8.38)$$

$$p_2 = p_{\text{в}}^* = \frac{Nf + \frac{1}{2}z_{\mathfrak{p}}^2 + z_{\mathfrak{p}}\sqrt{Nf(1-f) + \frac{1}{4}z_{\mathfrak{p}}^2}}{N + z_{\mathfrak{p}}^2}. \qquad (8.39)$$

Используем только что описанную методику для вычисления более точных интервальных оценок вероятности появления глагольных словоупотреблений в молдавских публицистических текстах.

Подставляя числовые значения в формулы (8.38) и (8.39), имеем

$$p_{\text{н}}^{*}=\frac{200\,000\cdot 0{,}1564+0{,}5\cdot 3{,}8416-1{,}96\sqrt{200\,000\cdot 0{,}1564\cdot 0{,}84+0{,}25\cdot 3{,}8416}}{200000+3{,}8416}+$$

$$+\frac{31281{,}92-1{,}96\sqrt{26388{,}77}}{200003{,}8416}=\frac{31281{,}92-318{,}34}{200003{,}8416}=\frac{30963{,}58}{200003{,}84}\approx 0{,}1548,$$

$$p_{\text{в}}^{*}=\frac{31281{,}92+318{,}34}{200003{,}84}=\frac{31600{,}26}{200003{,}84}\approx 0{,}1580.$$

Следовательно,

$$0{,}1548<p<0{,}1580.$$

Нетрудно заметить, что этот доверительный интервал несколько шире интервала, полученного по приближенной оценке. Это происходит потому, что улучшенная оценка дает меньшую погрешность по сравнению с приближенной. Расхождение между обеими оценками становится особенно заметным при малых выборках лингвистических единиц.

2. Интервальная оценка вероятности p для малых выборок. При малых N приближенные оценки вероятности p дают заметные ошибки, связанные с заменой неизвестного p опытным f, а также с переходом от дискретного биномиального к непрерывному нормальному распределению. Чтобы уменьшить эти ошибки, используются различные поправочные приемы.

Один из приемов состоит в том, что в левую и правую части двойного неравенства (8.32) вводят поправочный член $1/(2N)$, в связи с чем доверительный интервал p записывается следующим образом:

$$f-1/(2N)-z_{\mathfrak{p}}\sqrt{f(1-f)/N}<p<f+1/(2N)+z_{\mathfrak{p}}\sqrt{f(1-f)/N}. \tag{8.40}$$

Используем указанный прием для решения следующей лингвистической задачи. Из русского прозаического текста взята выборка в 50 словоупотреблений, в которой обнаружено 20 именных форм. Необходимо определить доверительные границы вероятности p при надежности в 95%.

Здесь $N=50$, $f=0{,}4$, $\mathfrak{p}=0{,}95$ и $z_{\mathfrak{p}}=1{,}96$. Подставив эти величины в формулу (8.40), получим

$$p_{\text{н}}^{*}=0{,}4-0{,}01-1{,}96\sqrt{0{,}4\,(1-0{,}4)/50}=0{,}2442\approx 0{,}24,$$

$$p_{\text{в}}^{*}=0{,}4+0{,}01+1{,}96\sqrt{0{,}4\,(1-0{,}4)/50}=0{,}5572\approx 0{,}56.$$

Другой прием заключается в том, что величина $z_{\mathfrak{p}}$ в формуле (8.32) заменяется на $t_{\mathfrak{p}}$ в распределении Стьюдента, а вместо

$\sqrt{f(1-f)/N}$ используют величину $\sqrt{f(1-f)/(N-1)}$. В этом случае доверительный интервал для p принимает вид

$$f-t_{\mathfrak{p},\nu}\sqrt{\frac{f(1-f)}{N-1}}<p<f+t_{\mathfrak{p},\nu}\sqrt{\frac{f(1-f)}{N-1}}. \tag{8.41}$$

Нижняя граница оценки равна

$$p_{\text{н}}^{*}=f-t_{\mathfrak{p},\nu}\sqrt{f(1-f)/(N-1)}, \tag{8.42}$$

а верхняя составляет

$$p_{\text{в}}^{*}=f+t_{\mathfrak{p},\nu}\sqrt{f(1-f)/(N-1)}. \tag{8.43}$$

Используем этот прием для определения доверительного интервала вероятности именных форм в только что рассмотренном примере.

Здесь $f=0{,}4$, $\mathfrak{p}=0{,}95$, $\nu=20-1=19$, $t_{0{,}95;19}=2{,}09$. Подставив эти значения в формулы (8.42) и (8.43), получим для нижней границы

$$p_{\text{н}}^{*}=0{,}40-2{,}09\sqrt{0{,}4\,(1-0{,}4)/49}=0{,}40-0{,}1463=0{,}2537\approx 0{,}25,$$

а для верхней границы

$$p_{\text{в}}^{*}=0{,}40+0{,}1463=0{,}5463\approx 0{,}55.$$

Легко заметить, что оба приема дают практически одни и те же оценки границ доверительного интервала.

Из сказанного не следует, что удовлетворительную интервальную оценку вероятности p можно получить при любом значении N. При очень малых объемах выборки даже самые сильные приемы оценки вероятности не дают удовлетворительного результата. В этом легко может убедиться читатель, взяв выборку в 5 словоупотреблений и задавшись целью определить при $\mathfrak{p}=0{,}95$ доверительный интервал вероятности появления существительных при условии, что их в этой выборке было всего 3 (т. е. 60%). При этом выясняется, что доверительный интервал здесь имеет вид $-0{,}08<p<1{,}28$, перекрывая таким образом весь возможный диапазон расположения истинной доли именных словоупотреблений в тексте. Это еще раз говорит о том, что построения «процентной» лингвистики, не учитывающие объема выборки и других важных понятий статистики, не всегда дают достаточное количество содержательной информации.

Все рассмотренные приемы интервальной оценки для вероятности p могут быть распространены и на логнормальное распределение.

3. Интервальная оценка вероятности редких лингвистических событий. До сих пор мы рассматривали случаи, когда вероятность лингвистической единицы достаточно велика. Такая ситуация типична для фонетико-фонологических и грамматических исследований. Однако при лексикологических исследованиях вероятности

словоформ, слов и словосочетаний обычно очень малы. Распределения их вероятностей чаще всего подчиняются распределению Пуассона.

Поэтому доверительный интервал вероятности лексических единиц следует устанавливать не на основе нормального приближения, которое может привести к значительным ошибкам, а путем применения пуассоновского приближения.

Для нахождения оценки p сначала нужно оценить параметр λ через число появлений события $F = Nf$, причем число всех испытаний N на оценку λ не влияет, оно может быть неизвестно, нужно лишь быть уверенным, что N велико. Распределение случайных величин F в распределении Пуассона оказывается тесно связанным с распределением χ^2 [7, с. 164]. Это и дает возможность получить доверительную оценку для λ.

Нижняя граница доверительного интервала пуассоновской вероятности задается выражением

$$p_{\text{н}}^* = \frac{1}{2N}\chi^2_{(1+\mathfrak{p})/2;\ \nu_1}, \tag{8.44}$$

где ν_1 — число степеней свободы, равное $2(F+1)$, а верхняя граница определяется из выражения

$$p_{\text{в}}^* = \frac{1}{2N}\chi^2_{(1-\mathfrak{p})/2;\ \nu_1}, \tag{8.45}$$

где число степеней свободы $\nu_2 = 2F$.

Решим следующую задачу. Из латышских научно-технических текстов взята выборка в 300 тыс. словоупотреблений, в которой слово aizgrieznis 'кран' встретилось 3 раза. Определить доверительные границы пуассоновой вероятности указанного слова при надежности $\mathfrak{p} = 0{,}95$.

Здесь $N = 300000$, $F = 3$, $\mathfrak{p} = 0{,}95$, $\nu_1 = 8$, $\nu_2 = 6$. С помощью табл. У (см. стр. 368) находим граничные значения:

$$\chi^2_{(1+\mathfrak{p})/2;\ \nu_1} = \chi^2_{0{,}975;\ 8} = 2{,}18,$$

$$\chi^2_{(1-\mathfrak{p})/2;\ \nu_2} = \chi^2_{0{,}025;\ 6} = 14{,}40.$$

Подставляя все необходимые значения в выражения (8.44) и (8.45), получаем значения

$$p_{\text{н}}^* = \frac{1}{2\cdot 300000}\cdot 2{,}18 = 0{,}0000036, \qquad p_{\text{в}}^* = \frac{1}{2\cdot 300000}\cdot 14{,}40 = 0{,}000024$$

нижней и верхней границ доверительного интервала, в котором заключена вероятность латышского слова aizgrieznis.

Методика интервальной оценки пуассоновой вероятности может быть применена также и к вероятностям, получаемым из лингвистических распределений Чебанова — Фукса и Фукса — Гачечиладзе.

§ 5. Оценка функции генерального распределения по данным лингво-статистического наблюдения

В предыдущих параграфах была рассмотрена первая задача, связанная с переходом от статистической модели текста к скрытой от прямого наблюдения вероятностной схеме его построения. Решение указанной задачи заключалось в оценке неизвестных параметров распределения генеральной лингвистической совокупности.

Теперь обратимся ко второй задаче, состоящей в том, чтобы по опытным данным оценить неизвестную функцию распределения непрерывной случайной лингвистической величины X. Решение этой задачи сводится к определению того доверительного интервала, в котором находятся значения теоретической функции, соответствующие определенным значениям аргумента. Этот доверительный интервал определяется по отдельным опытным данным без какого бы то ни было обращения к параметрам функции, которые остаются, как правило, нам неизвестными.

1. Доверительный интервал функции генерального распределения. При оценке неизвестной функции используются следующие рассуждения. Предположим, что в результате исследования текстовой выборки получена последовательность лингвистических единиц, упорядоченная в порядке возрастания номеров их извлечений из текста:

$$x_1, x_2, \ldots, x_N. \tag{8.46}$$

Величины x_i можно рассматривать в качестве значений некоторой конкретной случайной лингвистической величины X.

Последовательность (8.46) может быть описана с помощью интегральной функции эмпирического распределения накопленных частостей, соответствующих значениям x_i:

$$F_N(x_i) = \frac{F^*(x_i)}{N} = f_i^*.$$

Вместе с тем можно предположить, что имеется теоретический интегральный закон

$$F(x) = P(X < x_i) = p_i^*,$$

описывающий распределение значений рассматриваемой непрерывной случайной величины X в генеральной лингвистической совокупности. Вид этого закона нам неизвестен, и какой-либо дополнительной информацией о его свойствах, кроме свойства непрерывности, мы не располагаем.

Оценка теоретического закона $F(x)$ эмпирической функцией осуществляется, как и в случаях с неизвестными параметрами, при заданной надежности $\mathfrak{p}$ и точности ε. При этом доверительный интервал закона $F(x)$ имеет вид

$$F_N(x) - \varepsilon < F(x) < F_N(x) + \varepsilon.$$

Задачу определения доверительного интервала неизвестной теоретической функции $F(x)$ можно решать, исходя из интегральной предельной теоремы Муавра—Лапласа (см. гл. 6, § 3, п. 4).

Оценивая теоретический закон $F(x)$ с помощью эмпирической функции $F_N(x)$, мы будем считать, что значение $F_N(x)$ сходится по вероятности к $F(x)$, т. е.

$$F_N(x) = f^* \xrightarrow[N\to\infty]{\text{вер.}} p^* = F(x).$$

Тогда отношение между накопленной частостью и накопленной вероятностью при заданном x будет таким же, как и между относительной частотой и вероятностью в схеме Бернулли с N независимыми испытаниями. Следовательно, в качестве меры расхождения $F_N(x) - F(x)$ накопленной частости f^* от вероятности p^* можно взять среднее квадратическое отклонение

$$\sigma(f^*) = \sqrt{p^*(1-p^*)/N}.$$

Опираясь на указанную теорему, можно утверждать, что

$$\lim_{N\to\infty}\left\{-z < \frac{f^*-p^*}{\sqrt{p^*(1-p^*)/N}} < z\right\} = 2\Phi(z).$$

Иными словами, вероятность того, что абсолютная величина отклонения эмпирического распределения накопленных частот от теоретического интегрального закона будет меньше величины

$$z\sqrt{p^*(1-p^*)/N} = z\sigma(f^*),$$

при достаточно большом N приближается к $2\Phi(z)$:

$$\lim_{N\to\infty} P\{|f^*-p^*| < z\sigma(f^*)\} = 2\Phi(z).$$

Теперь можно утверждать с вероятностью $\mathfrak{p} = 2\Phi(z_\mathfrak{p})$, сколь угодно близкой к единице, что относительно каждой точки из последовательности (8.46) справедливо неравенство

$$-z_\mathfrak{p}\sigma(f^*) < f_i^* - p_i^* < z_\mathfrak{p}\sigma(f^*). \qquad (8.47)$$

Используем двойное неравенство (8.47) для построения доверительной полосы теоретической функции распределения $F(x)$. Для этого, задавшись вероятностью $\mathfrak{p}$ и определив из табл. VI на стр. 369 величину $z_\mathfrak{p}$, можно при известных $p^* = F(x)$ и N найти для каждого x_i величину $z_\mathfrak{p}\sigma(f^*)$. Откладывая эту величину вверх и вниз от точек $p_i^* = F(x_i)$, получим доверительную полосу (пунктирные линии на рис. 61). Если график эмпирической функции $F_N(x)$ (сплошная линия на рис. 61) всюду находится в пределах доверительной полосы, то вероятность события, состоящего в том, что для любого x отклонение $|f^* - p^*|$ меньше величины $z_\mathfrak{p}\sigma(f^*)$, равна $\mathfrak{p}$.

Положение доверительной полосы в значительной степени зависит от вида функции $F(x)$. Поэтому если $F_N(x)$ оказывается за

пределами полосы, то это может указывать на несоответствие теоретического и эмпирического распределений. В этом случае следует искать другую функцию $F(x)$, соответствующую эмпирической функции $F_N(x)$.

2. Построение доверительного интервала с помощью функции Колмогорова. Было бы неверным утверждать, исходя из всего сказанного, что выполнение неравенства

$$|f^* - p^*| < z_\mathfrak{p}\sigma(f^*) \qquad (8.48)$$

гарантируется с той же большой вероятностью $\mathfrak{p}$ одновременно для всех точек x_i.

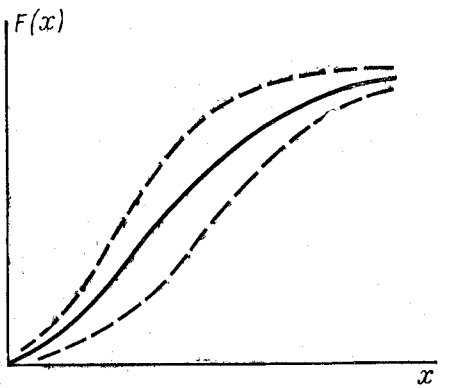


Рис. 61

В самом деле, несмотря на то что вероятность невыполнения неравенства (8.48) в определенной точке x_i очень мала, вероятность нарушения его хотя бы в одной точке x может оказаться достаточно большой. Отсюда следует, что проверка соответствия предполагаемого теоретического распределения опытному распределению на основании интегральной предельной теоремы не гарантирует от появления существенных отклонений $F_N(x_i) - F(x_i)$ для отдельных значений x_i. Можно показать, что при взятом нами значении $z_\mathfrak{p}\sigma(f^*) > 0$ максимальное расхождение эмпирических и теоретических значений

$$D_N = \max|F_N(x_i) - F(x_i)|$$

почти наверняка будет больше $z_\mathfrak{p}\sigma(f^*)$.

Одновременно возникает вопрос: что представляет собой функция распределения величины D_N и какова ее доверительная оценка? Ответ на этот вопрос дает **предельная теорема Колмогорова**, которая формулируется так: *если функция $F(x)$ непрерывна, то закон распределения*

$$P\{\max|F_N(x) - F(x)| < \lambda/\sqrt{N}\} = k_N(\lambda) \qquad (8.49)$$

не зависит от $F(x)$ и при $N \to \infty$ имеет пределом $K(\lambda)$, представляющую собой сумму бесконечного ряда:

$$K(\lambda) = \sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2\lambda^2}.$$

Значения функции Колмогорова $K(\lambda)$ приведены в табл. VII на стр. 369, 370.

При больших N выражение (8.49) можно переписать следующим образом:

$$P\{F_N(x)-\lambda/\sqrt{N} < F(x) < F_N(x)+\lambda/\sqrt{N}\} \approx K(\lambda),$$

где $\lambda/\sqrt{N}=\varepsilon$ можно считать величиной погрешности оценки закона $F(x)$, $K(\lambda)=\mathfrak{p}$ — надежностью, а $1-K(\lambda)=\mathfrak{q}$ — уровнем значимости (табл. 8.2).

Таблица 8.2

Надежность $K(\lambda)=\mathfrak{p}$	0,50	0,90	0,95	0,99	0,999
Уровень значимости $1-K(\lambda)=1-\mathfrak{p}=\mathfrak{q}$	0,50	0,10	0,05	0,01	0,001
$\lambda_{\mathfrak{p}}$	0,828	1,224	1,358	1,627	1,950

Если обратиться к геометрической интерпретации, то теорема Колмогорова утверждает с вероятностью $K(\lambda)$, что график неизвестной теоретической функции $F(x)$ целиком находится внутри полосы $C_K(\lambda)$, ограниченной кривыми $\underline{f}_i^* = F_N(x_i) - \varepsilon$ и $\overline{f}_i^* = F_N(x_i) + \varepsilon$ (рис. 62). Эти кривые получаются путем смещения вниз и вверх вдоль оси ординат эмпирических значений $F_N(x_i)$ на величину точности ε.

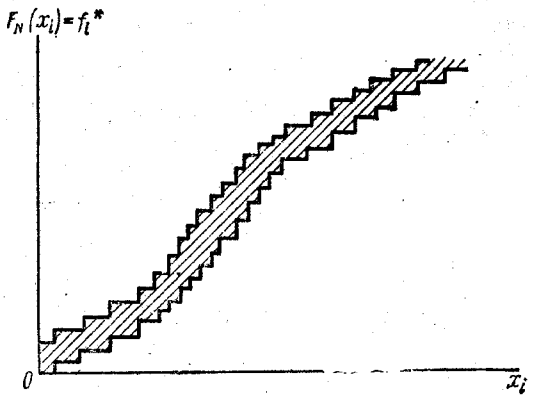


Рис. 62

Применение функции Колмогорова предполагает, во-первых, что теоретическая функция $F(x)$ непрерывна; во-вторых, что эмпирическая функция $F_N(x)$ построена по не сгруппированным в интервалы значениям случайной величины X. В практических приложениях допускается использование малых интервалов группировки, но получаемая оценка в этом случае действует с некоторым приближением.

Построение доверительной полосы для неизвестного теоретического закона распределения $F(x)$ с помощью функции Колмогорова осуществляется по следующей схеме.

1. Вычисляют значения эмпирической функции $F_N(x)$, представляющие собой накопленные частости

$$f_i^* = F_i^*/N. \qquad (8.50)$$

2. По заданному значению надежности $K(\lambda)=\mathfrak{p}$ с помощью табл. VII определяют значение $\lambda_{\mathfrak{p}}$ (обычно используют $\mathfrak{p}=0{,}90$, или 0,95, или 0,99).

3. С помощью равенства

$$\varepsilon_{\mathfrak{p}} = \lambda_{\mathfrak{p}}/\sqrt{N} \qquad (8.51)$$

определяют величину погрешности.

4. С помощью формул

$$\underline{f}_i^* = f_i^* - \varepsilon_{\mathfrak{p}}, \qquad (8.52)$$

$$\overline{f}_i^* = f_i^* + \varepsilon_{\mathfrak{p}} \qquad (8.53)$$

вычисляют границы доверительного интервала закона распределения $F(x)$ относительно точек x_i $(i=1, 2, \ldots, N)$.

Замечание. Если окажется, что $\underline{f}_i^* < 0$, а $\overline{f}_i^* > 1$, то следует считать $\underline{f}_i^* = 0$, а $\overline{f}_i^* = 1$.

5. На график наносят значения $\underline{f}_i^*$, $\overline{f}_i^*$ и таким образом определяют границы доверительной полосы (если имела место интервальная группировка данных, то графики $\underline{f}_i^*$ и $\overline{f}_i^*$ носят ступенчатый характер). Полученная полоса с вероятностью $\mathfrak{p}$ покрывает график теоретической функции $F(x)$.

Теперь, пользуясь данными о распределении длин китайских слогов (см. гл. 7, § 2, п. 2), дадим оценку неизвестного теоретического интегрального закона этого распределения (табл. 8.3).

Таблица 8.3

x_i	F_i	F_i^*	$F_N(x_i)=f_i^*=F_i^*/N$	$\underline{f}_i^*=f_i^*-\varepsilon_{\mathfrak{p}}$	$\overline{f}_i^*=f_i^*+\varepsilon_{\mathfrak{p}}$
(1)	(2)	(3)	(4)	(5)	(6)
50	2	2	0,0133	0	0,1246
70	7	9	0,0600	0	0,1713
90	7	16	0,1667	0,0554	0,2780
110	29	45	0,3000	0,1887	0,4113
130	29	74	0,4933	0,3820	0,6046
150	24	98	0,6533	0,5420	0,7646
170	17	115	0,7667	0,6554	0,8780
190	11	126	0,8400	0,7287	0,9513
210	12	138	0,9200	0,8087	1,0000
230	7	145	0,9667	0,8554	1,0000
250	1	146	0,9733	0,8620	1,0000
270	2	148	0,9867	0,8754	1,0000
290	1	149	0,9933	0,8820	1,0000
310	1	150	1,0000	0,8887	1,0000
	$N=150$				

Воспользуемся схемой построения доверительной полосы.

1. Приняв в качестве величин x_i середины интервальных значений длин китайских слогов (ср. табл. 7.8 на стр. 226), запишем в столбце (2) табл. 8.3 количество F_i слогов, длина которых соответствует данному интервалу, а в столбце (3) — абсолютные накопленные частоты F_i^* для каждого x_i. Пользуясь равенством (8.50) и учитывая, что общее число измеренных слогов N равно 150, в столбце (4) получим величины f_i^*, выступающие в качестве значений эмпирической функции $F_N(x)$.

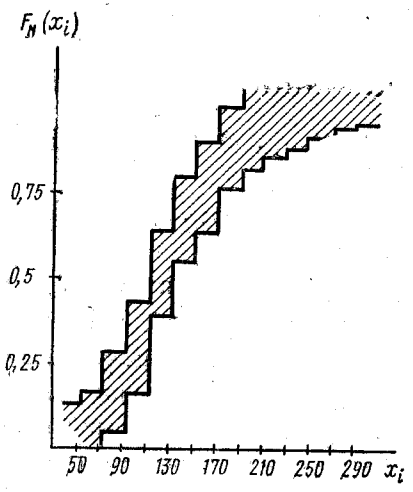


Рис. 63

2. Пусть надежность $\mathfrak{p}=0{,}95$, тогда по табл. VII получаем $\lambda_{\mathfrak{p}} = 1{,}358$.

3. На основании равенства (8.51) находим, что

$$\varepsilon_{\mathfrak{p}} = 1{,}358/\sqrt{150} = 1{,}358/12{,}206 \approx \approx 0{,}1113.$$

4. Пользуясь формулами (8.52) и (8.53), с учетом замечания, получаем значения нижней ($\underline{f}_i^*$) и верхней ($\bar{f}_i^*$) границ доверительной полосы теоретической функции распределения $F(x)$. Эти значения помещены в столбцах (5) и (6) табл. 8.3.

5. Нанеся на график значения $\underline{f}_i^*$ и $\bar{f}_i^*$, получаем графическое изображение доверительной полосы (рис. 63).

§ 6. Определение достаточности объема выборки

Ценность всякого лингвистического исследования измеряется степенью достоверности его выводов. Само собой разумеется, что лучшим средством оценки этой достоверности является проверка полученных выводов на практике. Но такая проверка может быть осуществлена в течение длительного времени после завершения самого исследования. Между тем лингвисту нужны приемы, с помощью которых можно было бы уже при постановке эксперимента прогнозировать достоверность получаемых результатов. Наиболее простым приемом такого прогнозирования является определение того минимального объема выборки, при котором получается заслуживающая доверия лингво-статистическая информация. Необходимый объем выборки можно определить, опираясь на оценку одного из параметров распределения, которое должно быть близким к нормальному, а также исходя из заранее установленной степени точности или относительной ошибки и надежности наших суждений.

Рассмотрим несколько вариантов определения достаточного объема лингвистической выборки.

1. Определение минимально достаточного объема выборки в грамматических и фонетико-фонологических исследованиях. Учитывая соотношение (6.129), можно утверждать, что величина максимально допустимой абсолютной ошибки равна

$$\varepsilon_{\max} = z_{\mathfrak{p}} \sqrt{p(1-p)/N} \qquad (8.54)$$

ср. с формулой (8.36)] при условии, что $|f-p| \leqslant \varepsilon$ [ср. с (6.115)].

Предположим, что величина абсолютной ошибки ε задана и заранее определена надежность $\mathfrak{p}$. Однако это не дает нам возможности вычислить величину выборки N, поскольку неизвестна вероятность p той лингвистической единицы, относительно которой определяется достаточность объема выборки. Чтобы оценить величину p, пользуются той частостью f лингвистической единицы, которая либо получена ранее в аналогичных условиях, либо извлечена из небольшого предварительного выборочного наблюдения. Заменяя p в выражении (8.54) на ее оценку f, получаем

$$\varepsilon = z_{\mathfrak{p}} \sqrt{f(1-f)/N}, \qquad (8.55)$$

откуда приходим к формуле, указывающей минимально достаточный в данных условиях объем выборки:

$$N = z_{\mathfrak{p}}^2 f(1-f)/\varepsilon^2. \qquad (8.56)$$

[ср. с (6.130)]. В тех случаях, когда дана не абсолютная, а относительная ошибка

$$\delta = \varepsilon/p \approx \varepsilon/f = z_{\mathfrak{p}} \sqrt{(1-f)/(Nf)} \qquad (8.57)$$

выражение (8.56) принимает вид

$$N = z_{\mathfrak{p}}^2 (1-f)/(\delta^2 f). \qquad (8.58)$$

Описанную процедуру целесообразно использовать при исследовании употребительности грамматических, фонетических и фонологических единиц, которые обычно дают нормальное распределение и вероятность которых не очень мала.

Например, по данным предварительного эксперимента относительная частота употребления мягких согласных фонем в украинском драматургическом тексте равна 0,0828 [34, с. 51]. Необходимо при заранее заданных максимальной абсолютной ошибке $\varepsilon = 0{,}0022$ и надежности $\mathfrak{p} = 0{,}95$ определить минимально достаточный объем выборки N для получения достоверных сведений об употребительности мягких согласных фонем в украинских драматургических текстах.

Здесь $f = 0{,}0828$, $1-f = 0{,}9172$, $\varepsilon = 0{,}0022$; по заданной величине $\mathfrak{p}$ из табл. VI на стр. 369 находим $z_{\mathfrak{p}} = 1{,}96$. Подставляя все эти данные в равенство (8.56), получаем

$$N = 1{,}96^2 \cdot 0{,}0828 \cdot 0{,}9172/0{,}0022^2 = 60284 \approx 60 \text{ тыс. фонем.}$$

Если вместо ε была бы задана величина δ, которая в данном случае равна 0,027, то для определения N следует воспользоваться

равенством (8.58); в конечном итоге мы получаем аналогичный результат:

$$N = 1{,}96^2 \cdot 0{,}9172/(0{,}027^2 \cdot 0{,}0828) = 58374 \approx 60 \text{ тыс. фонем.}$$

2. Определение минимально достаточного объема выборки в лексикологических исследованиях. В ходе лексикологических исследований минимальный объем выборки вычисляется с помощью приближенных равенств

$$N \approx z_p^2 f/\varepsilon^2 \qquad (8.59)$$

и

$$N \approx z_p^2/(\delta^2 f), \qquad (8.60)$$

которые легко выводятся из выражений (8.56) и (8.58), если принять во внимание, что частости f лексических единиц обычно очень малы и, следовательно, разности $1 - f$ близки к единице.

Рассмотрим в связи с этим следующую лингвистическую задачу. Относительная частота словосочетания split cylinder 'разрезной цилиндр' в английских текстах по строительным материалам составляет 0,000175 [33, с. 410]. Необходимо определить минимально достаточный объем выборки, удовлетворяющей надежности 0,95 и 33%-ной относительной ошибке.

Здесь $f = 0{,}000175$, $\delta = 0{,}33$ и $z_p = 1{,}96$ (при $p = 0{,}95$); поскольку f очень мало, для расчета величины N можно использовать формулу (8.60):

$$N = 1{,}96^2/(0{,}33^2 \cdot 0{,}000175) \approx 200000.$$

Таким образом, для получения достоверных результатов при статистическом исследовании лексических единиц, обладающих частостью не менее 0,000175, при заданной 33%-ной относительной ошибке и 95%-ной надежности нужна выборка объемом не менее 200 тыс. словоупотреблений.

В столбцах (4) — (6) табл. 8.4 показаны минимальные объемы выборки, необходимые для того, чтобы получить с надежностью p и относительной ошибкой δ (соответственно с абсолютной ошибкой ε) достоверные данные об употребительности лингвистических единиц, имеющих частоту выше заданного f.

Минимальные выборки, необходимые для получения достоверных данных об употребительности лингвистических единиц, имеющих частоту выше заданного f, если f достаточно мало, показаны в столбцах (7) — (9) табл. 8.4.

3. Определение достаточности объема выборки с учетом рассеяния лингвистического признака. В предыдущих разделах объем выборки мы определяли, исходя из оценки только одного параметра распределения — вероятности p. Однако в лингвистических и инженерно-лингвистических исследованиях нередко встречаются задачи, которые требуют, чтобы при определении минимального объема выборки были учтены средняя частота $\overline{F}$, оценивающая матема-

Таблица 8.4

Необходимые объемы выборки при известных δ (или ε), p (z_p) и f

(p оценено через f)

f	ε	δ	$N = \frac{z_p^2(1-p)}{p\delta^2} \approx \frac{z_p^2(1-f)}{f\delta^2}$			$N = \frac{z_p^2}{f\delta^2}$			$N = \frac{9}{\frac{p(1-p)}{(1-f)}} \approx \frac{9}{f}$
			$z_p = 1{,}65$	$z_p = 1{,}96$	$z_p = 2{,}58$	$z_p = 1{,}65$	$z_p = 1{,}96$	$z_p = 2{,}58$	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
0,0001	0,00001	0,1	2720000	3840000	6650000	2720000	3840000	6660000	90000
	0,00002	0,2	680000	960000	1660000	680000	960000	1660000	
	0,00003	0,3	300000	430000	740000	300000	430000	740000	
0,001	0,0001	0,1	272000	384000	665000	272000	384000	666000	9000
	0,0002	0,2	68000	96000	166000	68000	96000	166000	
	0,0003	0,3	30000	43000	74000	30000	43000	74000	
0,01	0,001	0,1	27000	38000	66000	27200	38400	66600	910
	0,002	0,2	6700	9500	16500	6800	9600	16600	
	0,003	0,3	3000	4200	7200	3000	4300	7400	
0,05	0,005	0,1	5200	7300	12600	5450	7700	13300	190
	0,010	0,2	1300	1800	3200	1360	1920	3300	
	0,015	0,3	575	810	1400	600	850	1580	
0,10	0,01	0,1	2450	3460	6000	2720	3840	6600	100
	0,02	0,2	610	860	1500	680	960	1660	
	0,03	0,3	270	380	670	300	430	740	
0,20	0,02	0,1	1100	1540	2660	1360	1920	3330	60
	0,04	0,2	270	380	670	340	480	830	
	0,06	0,3	120	170	300	150	210	370	
0,50	0,05	0,1	270	380	670	540	770	1330	40
	0,10	0,2	70	100	170	140	190	330	
	0,15	0,3	30	40	70	60	80	150	

тическое ожидание $M(X)$ этого признака, и стандарт s, оценивающий среднее квадратическое отклонение $\sigma = D(X)$ (величины δ и p считаются при этом заданными).

Величина выборки определяется здесь исходя из следующих соображений.

На основании центральной предельной теоремы Ляпунова (см. гл. 6, § 4, п. 4), можно утверждать, что погрешность $|F - \overline{F}|$ распределена нормально. Максимальная величина этой погрешности $\varepsilon_{max} = \max |F - \overline{F}|$ может быть определена как

$$\varepsilon_{max} = \overline{F}\delta = z_p \delta / \sqrt{N} \approx z_p \delta / \sqrt{N}.$$

Отсюда

$$N = z_p^2 s^2 / (\overline{F}^2 \delta^2). \qquad (8.61)$$

Построение двуязычных и одноязычных машинных словарей связано с разного вида сжатием лексической информации. При этом важно иметь очень точные сведения о средней длине словарного слова или словоформы. Эти сведения можно получить, обследуя не тексты, а одноязычные или двуязычные словари разного объема.

Так, например, предварительное обследование русского технического словника показало, что средняя длина словоформы равна здесь 8,68 буквы при достаточно высоком рассеянии численных значений ее длины: стандарт здесь составляет 3,12 буквы. Нужно определить минимально необходимый объем словаря L, из которого можно получить при $\delta = 0{,}01$ и $p = 0{,}95$ достоверные данные о средней длине словоформы.

Здесь $\overline{F} = 8{,}68$, $s = 3{,}12$, $\delta = 0{,}01$, $p = 0{,}95$ ($z_p = 1{,}96$). Подставляя эти данные в формулу (8.61), получаем

$$L = 1{,}96^2 \cdot 3{,}12^2 / (8{,}68^2 \cdot 0{,}01^2) = 4970 \text{ словоформ.}$$

В работе [7, с. 184] показано, что в тех случаях, когда дисперсия $D(X) = \sigma^2 = Np(1-p)$ больше или равна 9, случайная величина распределена нормально. Отсюда можно утверждать, что величину минимально необходимой выборки для определенного значения p, оцениваемого частостью f, можно установить из неравенства

$$N \geqslant \frac{9}{f(1-f)}. \qquad (8.62)$$

Из столбца (10) табл. 8.4 видно, что формула (8.62) указывает объемы выборок, в несколько раз меньшие чем минимально необходимые объемы выборок, рассчитанные по формулам (8.56) — (8.60). Однако эта оценка выборки не гарантирует какой-либо надежности p и не фиксирует какую-либо определенную величину относительной ошибки δ.

4. Оценка точности результатов лингво-статистических исследований. В лингвистических исследованиях нередко приходится оценивать точность получаемых статистических результатов. Для

оценки точности можно использовать величины относительной ошибки δ и погрешности (абсолютной ошибки) ε, которые можно найти по приведенным выше формулам. При этом нужно помнить, что степень точности находится в обратной зависимости по отношению к величинам δ и ε: чем выше точность лингво-статистического измерения, тем меньше должны быть величины относительной ошибки или погрешности.

Степень точности лингво-статистических измерений связана также с объемом выборки и надежностью.

Например, при исследовании устного украинского драматургического текста длиной в 60 тыс. фонем обнаружено, что относительная частота мягких согласных фонем равна 0,0828. Необходимо определить абсолютную ошибку ε и относительную ошибку δ при получении указанного результата, если надежность $p = 0{,}95$ ($z_p = 1{,}96$).

Пользуясь формулами (8.54) и (8.57), получаем

$$\varepsilon = 1{,}96 \sqrt{0{,}0828 \cdot 0{,}9172 / 60000} \approx 0{,}0022,$$

$$\delta = 1{,}96 \sqrt{0{,}9172 / (60000 \cdot 0{,}0828)} \approx 0{,}027 = 2{,}7\ \%.$$

Из выражений (8.56) и (8.58) видно, что величины погрешности и относительной ошибки находятся в обратной зависимости по отношению к объему выборки. Чем больше объем лингво-статистических наблюдений, тем меньше погрешность и относительная ошибка и тем выше точность наших лингво-статистических результатов. В только что рассмотренном примере получена довольно высокая точность наблюдений (всего 2,7% относительной ошибки). Однако эта точность была оплачена довольно дорогой ценой: пришлось просчитать текст длиной в 60 тыс. фонемоупотреблений (т. е. около 30 страниц среднего объема). Можно было бы при желании уменьшить относительную ошибку в 10 раз, доведя ее до 0,27%, но тогда выборку пришлось бы увеличить в 100 раз, просчитав 2400 страниц фонетической записи украинского текста. В связи с этим возникает вопрос: каков же разумный предел повышения точности лингво-статистических измерений?

Выбор уровня точности, так же как и выбор надежности наших суждений, зависит от теоретических и практических приложений той дисциплины, которая использует описываемые статистические приемы. Если для авиационной и ракетной техники относительная ошибка измерения в 2,7% может рассматриваться как предельная, то для лингвистических исследований такая точность оказывается излишней и приводит к неразумному увеличению объема выборки и, таким образом, к неоправданному расходованию сил исследователя на механическую нетворческую работу. В настоящее время принято считать, что в фонетико-фонологических и грамматических исследованиях относительная ошибка не должна превышать 20%, а при анализе лексики и фразеологии относительная ошибка может достигать 30 — 35% [7, с. 189].

5. Определение достаточности объема выборки по заданной покрываемости текста. Первая задача, с которой сталкивается лингвист, приступающий к составлению частотного списка букв, звуков, фонем, слогов, морфем, слов, словоформ, словосочетаний или грамматических форм,—это, как уже говорилось, определение минимально достаточного объема текста.

Существуют различные подходы к решению этой задачи. Иногда требование достаточности заменяется «критерием» осуществимости: в этом случае определяется такой объем выборки, который можно произвести в отведенный для этого срок.

При определении длины обследуемого текста можно исходить из требования, согласно которому в частотный список попала бы заранее заданная доля разных слов (соответственно букв, слогов и т. д.), образующих данный текст.

Однако чаще всего объем выборки определяется исходя из требований к покрываемости наугад взятого текста наиболее частыми единицами составляемого списка. При этом предполагается, что частоты и частости этих единиц будут достоверными, т. е. определены в пределах допустимой относительной ошибки и с заданной надежностью. Чтобы решить эту задачу, необходимо сначала выяснить, чему равна частость той единицы частотного списка, которая выступает в качестве нижней границы массива, дающего заданную покрываемость текста. Будем обозначать эту относительную частоту термином *граничная частость* ($f_{гр}$). Если в частотном списке указываются относительные накопленные частоты отдельных единиц, то граничная частость находится на той накопленной частоте, которая соответствует заданной покрываемости текста ($f\dot{c}$).

При определении объема выборки, обеспечивающей заданную покрываемость текста, кроме величин $f_{гр}$ и $f\dot{c}$, приходится учитывать также относительную ошибку δ и надежность p. Если фиксировать для $f_{гр}$ относительную ошибку δ_{max}, то для частостей $f_i > f_{гр}$ имеет место неравенство $\delta_i < \delta_{max}$. Таким образом, можно утверждать, что для массива единиц частотного списка, обладающего покрываемостью C, относительная ошибка частости не превосходит заданной величины δ_{max}. Итак, задав относительную ошибку δ_{max}, надежность p и соответствующую ей величину z_p, а также определив опытным путем по покрываемости $C = f\dot{c} \cdot 100\%$ граничную частость $f_{гр}$, мы можем согласно формуле (8.60) вычислить минимально необходимый объем выборки; в тех случаях, когда $f_{гр}$ очень мало, он равен

$$N(C) = \frac{z_p^2}{\delta_{max}^2 f_{гр}}. \qquad (8.63)$$

Для тех случаев когда $f_{гр}$ заметно отличается от нуля, исходя из (8.58), получаем

$$N(C) = \frac{z_p^2 (1 - f_{гр})}{\delta_{max}^2 f_{гр}}. \qquad (8.64)$$

Проиллюстрируем описанную процедуру двумя примерами.

1. Пусть необходимо определить минимальный объем казахского текста для получения достоверных статистических данных относительно массива наиболее частых словоформ, покрывающих 80% текста. Относительная ошибка не должна превышать 30%, надежность $p = 0{,}95$.

По частотному словарю казахской публицистики находим значение накопленной частоты $f\dot{c} = 0{,}8$, соответствующее заданной покрываемости, это значение имеет словоформа с порядковым номером $i = 5503$ и $f_{гр} = 0{,}00004$. Учитывая, что $\delta_{max} = 0{,}3$, $z_{0,95} = 1{,}96$, а $f_{гр}$ достаточно мало, используем для определения минимального объема выборки выражение (8.63). Тогда получим

$$N_{C=80\%} = 1{,}96^2/(0{,}33^2 \cdot 0{,}00004) \approx 880\,000 \text{ словоупотреблений.}$$

2. Пусть требуется определить минимально необходимый объем русского разговорного текста для получения достоверных статистических данных с относительной ошибкой в 30% и надежностью $p = 0{,}95$ о массиве наиболее частых служебных слов. Этот массив дает покрываемость $C = 17\%$.

По данным пробного частотного словаря русских разговорных текстов находим, что $C = 17\%$ соответствует $f_{гр} = 0{,}013$; учитывая также, что $\delta_{max} = 0{,}3$, а $z_p = 1{,}96$, получаем

$$N_{C=17\%} = \frac{1{,}92^2 (1 - 0{,}01)}{0{,}3^2 \cdot 0{,}01} = \frac{3{,}84 \cdot 0{,}99}{0{,}09 \cdot 0{,}01} = 4224 \text{ словоупотребления.}$$

ГЛАВА 9

ИССЛЕДОВАНИЕ ВЕРОЯТНОСТНЫХ СВОЙСТВ ЯЗЫКА И СТАТИСТИКИ ТЕКСТА С ПОМОЩЬЮ МЕТОДА ГИПОТЕЗ

§ 1. Элементы теории статистических гипотез

1. Статистические и нестатистические гипотезы в лингвистике. Рассмотренные в предыдущей главе процедуры перехода от текстовой статистической модели (вариационного ряда) к генеральному распределению давали возможность обнаружить количественные параметры нормы, однако лингвистику интересуют не только количественные характеристики языка и речи, но и в первую очередь их качественные признаки.

Среди различных научных методов, с помощью которых могут быть обнаружены, описаны и количественно оценены качественные лингвистические явления, важное место занимает метод статистической проверки лингвистических гипотез.

Прогресс каждой науки, в том числе и языкознания, связан с выдвижением и проверкой гипотез. Использующиеся в науке гипотезы очень разнообразны как по способам их формулировки, так и по методам их проверки. Среди всего разнообразия этих гипотез особое место занимают *статистические гипотезы*, т. е. такие гипотезы, которые формулируются либо относительно вида распределения случайной величины, либо относительно параметров распределения, либо относительно ранговой упорядоченности значений случайной величины. Будучи сформулированными относительно вероятностно-статистических и ранговых величин, эти гипотезы могут проверяться и оцениваться с помощью разного вида статистических приемов и критериев. По результатам проверки и оценки статистических гипотез мы получаем возможность делать качественные лингвистические выводы.

Рассмотрим несколько статистических и нестатистических гипотез.

Статистическими можно считать гипотезы о том, что служебные слова в тексте имеют нормальное (или же пуассоновское) распределение, а употребление большинства знаменательных слов, особенно слов терминологического значения, не подчиняется этим видам распределения. Действительно, обследуя выборочные значения средних частот интересующих нас слов по различным сериям, мы можем оценить, насколько близко рассматриваемое эмпирическое распределение к теоретическому распределению Пуассона или же к нормальному распределению.

К статистическим гипотезам можно отнести предположение о том, что употребление служебных и многих знаменательных слов в немецких публицистических текстах из ГДР и ФРГ подчиняется одним и тем же вероятностным нормам. Это предположение можно проверить путем сравнения частостей соответствующих слов в газетных текстах ГДР и западногерманских публицистических текстах. Также статистическим может считаться предположение о том,

что средняя длина слова, измеренная в буквах латинского или кириллического алфавитов относительно научно-технического и делового стиля, во всех языках мира равна 5,5 буквам, а дисперсия составляет 1,2 буквы. Эти гипотезы рассматривают значения параметров в генеральных лингвистических распределениях, и их можно проверить, взяв случайные лингвистические выборки. Напротив, к статистическим гипотезам нельзя отнести предположение Тура Хейердала о южноамериканском происхождении населения острова Пасхи, равно как и альтернативную гипотезу Те Ранги Хироа (П. Бака) об этнической связи жителей этого острова с древним населением Юго-Восточной Азии. Нестатистическими являются гипотезы о иберо-кавказском происхождении баскского языка или предположение С. Девиса о том, что фестский диск написан похеттски. Все эти гипотезы сформулированы и проверяются с помощью неколичественного научного аппарата, они не рассматривают ни вид распределения, ни величины параметров, ни ранговые последовательности и поэтому не могут быть проверены статистическим путем. С другой стороны, далеко не все оперирующие количественными данными научные концепции могут рассматриваться в качестве статистических гипотез. Примером могут служить глоттохронологические гипотезы (см. гл. 2, § 3, п. 1 и 2), которые хотя и используют количественные данные, но не могут быть сформулированы относительно какого-либо вида распределения, параметра или ранговой последовательности.

2. Нулевая и альтернативная гипотезы. Ошибки первого и второго рода. Утверждение о предполагаемом виде распределения, ранговой последовательности или параметре формулируется в виде *нулевой* (или основной) *гипотезы* H_0, которой противопоставляется другая — *альтернативная гипотеза* H_1. Например, предположение о том, что служебные слова в тексте имеют пуассоновское (или же нормальное) распределение, можно считать нулевой гипотезой H_0. Утверждение же о том, что указанные слова не подчиняются пуассоновскому (или нормальному) распределению, в этом случае выступает в качестве альтернативной гипотезы H_1.

Альтернативных гипотез может быть несколько. Например, если считать предположение о том, что средняя длина слова во всех языках мира равна 5,5 буквы, то в качестве альтернативных гипотез H_1, H_2, ..., H_i, ..., H_n, выступают предположения о том, что средняя длина слова равна 5, 6, ..., 20 и т. д. буквам.

В дальнейшем мы будем рассматривать лишь такие случаи, когда речь идет о нулевой гипотезе H_0 и одной альтернативной гипотезе H_1. Последняя может объединять несколько альтернативных гипотез, выступая в качестве отрицания нулевой гипотезы, т. е. $H_1 = \overline{H}_0$.

Проверка лингво-статистической гипотезы всегда осуществляется на случайной выборке. Поскольку выборка конечна, она не может идеально точно отразить распределение в генеральной лингвистической совокупности. Вместе с тем всегда существует риск сформировать такую «неудачную» выборку, которая дала бы совершенно

ложную информацию о положении дел в генеральной лингвистической совокупности (вряд ли следовало бы проверять лингво-статистические гипотезы, относящиеся к мертвым языкам во всем их стилевом разнообразии, на выборках, составленных из культовых текстов или надгробных надписей). Короче говоря, при проверке лингво-статистической гипотезы всегда есть шанс прийти к ложному решению.

В результате проверки лингвистической гипотезы с помощью того или иного статистического критерия возникает одна из следующих четырех ситуаций:

1) нулевая гипотеза H_0 принимается, и она верна (соответственно отвергается ложная альтернативная гипотеза H_1);

2) нулевая гипотеза H_0 отвергается, и она ложна (соответственно принимается верная альтернативная гипотеза H_1);

3) нулевая гипотеза H_0 отвергается, хотя она и верна (соответственно принимается ложная гипотеза H_1);

4) нулевая гипотеза H_0 принимается, хотя она и ложна (соответственно отвергается правильная альтернативная гипотеза H_1).

Первые две ситуации представляют собой правильные решения, а две последние — ошибочные решения. При этом третье решение, состоящее в отвержении правильной гипотезы H_0, дает *ошибку первого рода*, в то время как четвертое решение, заключающееся в принятии нулевой гипотезы H_0, хотя она ложна, представляет собой *ошибку второго рода* (табл. 9.1).

Таблица 9.1

	Гипотеза H_0 верна	Гипотеза H_0 не верна
Гипотеза H_0 отвергается	Ошибка первого рода	Правильное решение
Гипотеза H_0 принимается	Правильное решение	Ошибка второго рода

Может показаться, что выбор одной из двух возможных гипотез в качестве нулевой, т. е. основной, а другой — в качестве альтернативной является совершенно произвольным и определяется соглашением. Этим соглашением определяется и то, какое из неправильных решений считать ошибкой первого рода, а какое — ошибкой второго рода.

Предположим, например, что алтайскую гипотезу о генетическом родстве тюркских, монгольских, тунгусо-маньчжурских языков удалось сформулировать таким образом, что ее можно проверить с помощью некоторого статистического критерия. В этом случае совершенно безразлично, будем ли мы считать нулевой гипотезой предположение о том, что эти языки родственны, или объявим в качестве гипотезы H_0 утверждение о том, что эти языки не имеют генетического родства. В первом случае ошибка первого рода состоит в отрицании родства этих языков, хотя они в действительности восходят

к одному источнику. Во втором случае ошибкой первого рода служит утверждение о том, что алтайские языки родственны, хотя в действительности они неродственны. Аналогичным образом в первом случае ошибкой второго рода будет утверждение о родстве алтайских языков (принятие гипотезы H_0, в то время как языки эти не имеют генетического родства), а во втором случае ошибкой второго рода — утверждение о том, что языки не имеют генетического родства, хотя они и восходят к одному источнику.

Из приведенного примера видно, что с точки зрения широкой лингвистической общественности выбор нулевой гипотезы при решении алтайской проблемы является чисто условным, а различия между ошибками первого и второго рода не являются значимыми.

Однако встречаются ситуации, когда определение того, какую гипотезу считать основной (нулевой), небезразлично для исследователя, и в связи с этим ошибки первого и второго рода получают разную значимость.

Предположим, что имеется программа машинного перевода текстов военно-оперативной тематики с одного языка (этим языком может быть как язык противника, так и язык союзника) на другой. Техническое исполнение программы сомнений не вызывает, необходимо оценить лингвистический алгоритм. Лингвистическое качество программы экспериментально проверяется на малых порциях текста. Результаты эксперимента дают частные значения x_1, x_2, ..., x_N случайной величины X, в качестве которой выступает число неправильно переведенных фраз порции.

При такой постановке эксперимента мы можем сформулировать две лингво-статистические гипотезы. Первая (гипотеза A_1) утверждает, что программа пригодна и ее можно принять на вооружение. Вторая (гипотеза A_2) состоит в том, что программа непригодна для использования в войсках и ее следует вернуть на доработку.

Возьмем в качестве нулевой гипотезу A_1, тогда ошибка первого рода заключается в том, что пригодная программа будет направлена на доработку. Это приведет к потере времени и к дополнительным затратам средств, прежде чем будет выяснено, что программа в доработке не нуждается.

Если же принять в качестве основной гипотезы предположение о том, что программа непригодна (A_2), то ошибка первого рода состоит в принятии негодной программы, использование которой может привести к срыву военных операций.

Нетрудно понять, что ошибка первого рода, могущая возникнуть в том случае, когда $H_0 = A_2$, гораздо более серьезна, чем ошибка первого рода, которая может образоваться при проверке нулевой гипотезы $H_0 = A_1$.

Заметим попутно, что в ситуации, при которой $H_0 = A_2$, ошибкой второго рода является ошибка первого рода ситуации $H_0 = A_1$ (т. е. ошибка, состоящая в отвержении доброкачественной программы), и наоборот, ошибка первого рода ситуации $H_0 = A_2$ служит ошибкой второго рода для ситуации $H_0 = A_1$.

Только что рассмотренный пример показывает, что при испытании гипотез часто возникают такие ситуации, при которых избежать одной из двух возможных ошибок оказывается важнее, чем допустить другую. В этом случае ошибкой первого рода считается та из возможных ошибок, которую нам важнее избежать. Отсюда следует, что нулевой гипотезой следует считать то предположение, отвержение которого в том случае, когда оно является истинным, приводило бы к ошибке первого рода.

Возвращаясь к примеру с машинным переводом, нетрудно показать, что в качестве нулевой гипотезы следует принять предположение A_2 о том, что программа машинного перевода военно-оперативных текстов непригодна для использования в войсках. В этом случае мы должны будем стараться максимально уменьшить ошибку первого рода, состоящую в том, что в результате ошибочного отвержения гипотезы $H_0 = A_2$ на вооружение поступает недоброкачественная программа машинного перевода.

Разумеется, определение значимости ошибки, особенно в лингвистике, носит часто субъективный характер. Так, например, для убежденного алтаиста наиболее опасной ошибкой (ошибкой первого рода) служит отвержение нулевой гипотезы, состоящей в том, что алтайские языки восходят к одному источнику. Ошибка второго рода, состоящая в принятии алтайской гипотезы, хотя последняя неверна, будет, разумеется, менее опасна для алтаиста.

3. Проверка статистических гипотез. Проверка статистических гипотез опирается на такие фундаментальные понятия, как критическая область, уровень существенности (уровень значимости), критерий проверки, а также мощность критерия.

Начнем с разъяснения понятия *критической области*, используя при этом теоретико-множественные представления из аксиоматического построения теории вероятностей А. Н Колмогорова (см. гл. 5, § 3, п. 4). Пусть U обозначает множество (пространство) значений случайной величины X, а W — выборочное, наблюдаемое в опыте подмножество этих значений. Подмножество W состоит из возможных выборочных точек $e_1, e_2, \ldots, e_m$, каждая из которых представляет собой определенную совокупность наблюдаемых значений $x_1, x_2, \ldots, x_N$ случайной величины. Точки $e_1, e_2, \ldots, e_m$, в свою очередь, можно рассматривать как некоторые выборочные числовые значения некоторой статистической характеристики E, относительно которой выдвигаются гипотезы H_0 и H_1. Нулевая гипотеза H_0 проверяется с помощью некоторого статистического критерия С. Гипотеза эта отвергается лишь тогда, когда наблюдаемая характеристика, т. е. выборочная точка e_i, попадает внутрь определенной области подмножества W — в так называемую критическую область W_C. Если эта наблюдаемая характеристика попадает внутрь области $W - W_C$, называемой *областью приемлемости решений*, то гипотеза H_0 принимается (рис. 64).

Нетрудно понять, что выбор критерия проверки гипотезы эквивалентен выбору критической области, а задача проверки гипотезы является задачей выбора критической области. Основная цель испытания гипотезы состоит в том, чтобы уменьшить ошибку в принятии решения, причем избежать ошибки первого рода важнее, чем допустить ошибку второго рода. Поэтому статистический критерий, а следовательно и критическая область, должны быть выбраны таким образом, чтобы в случае справедливости нулевой гипотезы H_0 выборочные наблюдения e_i попадали бы в критическую область W_C как можно реже. С этой целью фиксируется произвольно малое число α, называемое *уровнем существенности*, и выдвигается требование, согласно которому вероятность ошибки первого рода при проверке истинной гипотезы H_0 не должна превосходить этого уровня.

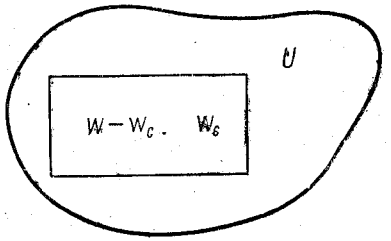


Рис. 64

Итак, между понятиями критической области, статистической характеристикой гипотезы и уровнем существенности имеет место следующее соотношение: критическая область W_C должна быть такова, что вероятность принятия статистической характеристикой E числового значения, попадающего в W_C, не превышает уровня существенности α (при этом предполагается, что гипотеза H_0 истинна).

Выбор уровня существенности определяется практическими соображениями, точнее, ожидаемыми последствиями ошибки первого рода. Чем серьезнее эти последствия, тем должен быть меньше уровень существенности. В лингвистике обычно используется $\alpha = 0{,}05$. Такой уровень существенности вполне приемлем, например, при статистическом испытании алтайской гипотезы. Однако при проверке ответственных инженерно-лингвистических гипотез этот уровень целесообразно значительно уменьшить: так, при проверке нулевой гипотезы о непригодности программы машинного перевода следовало бы взять уровень существенности, не превышающий 0,01. Здесь мы имели бы только один шанс из ста принять на вооружение непригодную программу.

При рассмотрении проблемы выбора уровня существенности в тех случаях, когда ответственность за ошибку первого рода достаточно велика, может возникнуть вопрос: нельзя ли настолько уменьшить уровень существенности, чтобы вероятность ошибки первого рода была бы близка к нулю?

Чтобы ответить на этот вопрос, предположим, что мы взяли $\alpha = 0$. Тогда независимо от результатов опыта нулевая гипотеза H_0 будет приниматься и в том случае, когда она верна, и в том случае, когда она ложна. Принимая нулевую гипотезу в тех случаях, когда она ложна, мы совершаем, как уже говорилось, ошибку второго рода. При бесконечном уменьшении α вероятность ошибки второго рода возрастает, что часто приводит к нежелательным результатам.

Например, проверяя качество программы машинного перевода, мы должны при любом результате испытания нулевой гипотезы, состоящей в предположении, что программа перевода непригодна, принимать эту гипотезу, отвергая альтернативную гипотезу о пригодности программы. Цена этого скептицизма будет достаточно велика: органы управления и связи вооруженных сил не получат ни одной программы машинного перевода военно-оперативных текстов.

Итак, стремясь к уменьшению вероятности ошибки первого рода, мы не должны забывать о том, что оно может привести к нежелательному увеличению вероятности β ошибки второго рода. Оптимальное соотношение вероятностей α и β достигается при удачном выборе критической области. Условие такого выбора состоит в следующем: *если вероятность того, что выборочное значение статистической характеристики e_t попадет в критическую область W_C при справедливости гипотезы H_0, составляет*

$$P\{e_t \in W_C/H_0\} = \alpha, \qquad (9.1)$$

то вероятность того, что это же значение e_t попадет в область W_C при неверной гипотезе H_0 и истинной гипотезе H_1, должна иметь максимальное значение, т. е.

$$P\{e_t \in W_C/H_1\} = \max. \qquad (9.2)$$

Тогда вероятность β ошибки второго рода, состоящей в том, что ошибочно отбрасывается верная гипотеза H_1, является минимальной.

Вся процедура проверки нулевой гипотезы осуществляется с помощью определенного критерия С, представляющего собой правило, которое устанавливает, при каких результатах случайной выборки мы имеем право принять нулевую гипотезу, а при каких — отвергнуть. Требование, выраженное равенством (9.2), называется *постулатом мощности критерия* С, мощности, которая измеряется вероятностью того, что не будет допущена ошибка второго рода.

§ 2. Гипотеза о лексической нормативности текста и ее проверка с помощью порядковых критериев

1. Порядковые и статистические критерии. Метод статистических гипотез особенно широко используется при определении нормативности лексических и грамматических единиц, а также при сопоставлении стилевых и жанровых разновидностей языка [32 а, с. 16 и сл.]; [33, с. 131 и сл.]. Этот прием может быть применен и для выявления лексико-грамматических особенностей стиля отдельных авторов.

Исследования этого типа предполагают количественное сопоставление двух лингвистических множеств (частотных списков, распределений частот слова или грамматической формы и т. п.), одно из которых может рассматриваться в качестве эталонного. Выдвигается предположение о том, что оба множества по своей

лингвистической природе идентичны и статистически однородны. Если это предположение рассматривать в качестве нулевой гипотезы H_0, то альтернативной гипотезой H_1 служит утверждение, что сравниваемые множества имеют разную лингвистическую природу.

Гипотеза H_0 проверяется, как уже говорилось, с помощью некоторого объективного критерия С, который должен количественно оценить степень (силу) сходства между сопоставляемыми лингвистическими множествами. Если оценка сходства попадает в область приемлемости решений $W - W_C$ (см. выше), то гипотеза об идентичности сравниваемых лингвистических множеств принимается.

Наиболее мощными критериями контроля лингвистических предположений являются *статистические критерии*, с помощью которых либо проверяется нулевая гипотеза применительно к значениям некоторых параметров распределения, либо оцениваются гипотезы о характере самих распределений при условии, что оценка строится на исследовании параметров распределения. Такие правила проверки называются *параметрическими критериями*. Более простыми с точки зрения используемого математического аппарата являются *непараметрические* статистические *критерии*, которые, используя лишь частости или частоты лингвистических единиц, не требуют знания параметров распределения. Непараметрические правила проверки гипотез менее эффективны, чем параметрические критерии. Бывает, что проверку предположения об идентичности двух лингвистических множеств нельзя провести с помощью указанных статистических критериев ввиду того, что нет надежных оценок параметров распределения, неизвестен его характер, а также отсутствуют достоверные значения (частоты) лингвистических величин. В этих случаях для проверки статистической гипотезы используются *порядковые критерии*, в которых применяются не сами значения случайных величин, а ранговая упорядоченность этих значений. Использование рангов вместо частот связано с потерей части статистической информации и влечет за собой снижение мощности критерия, увеличивая тем самым вероятность ошибки второго рода (ср. § 1, п. 2).

Несмотря на эти недостатки, порядковые критерии являются наиболее простыми и универсальными приемами оценки лингвистических гипотез.

Чаще всего эти критерии используются для проверки гипотезы об однородности двух текстовых выборок относительно заданного лингвистического признака или гипотезы об идентичности (нормативности) функционирования данного лингвистического признака в обеих выборках.

Идею проверки лингвистической гипотезы с помощью порядкового критерия можно сформулировать следующим образом.

Имеется лингвистический признак L, в роли которого может выступать конкретное слово, грамматическая категория и т. п. Взяты также две текстовые выборки N_1 и N_2, разбитые на n порций каж-

дая. В выборке N_1 лингвистический признак L может рассматриваться в качестве случайной величины X, имеющей значения $x_1, x_2, \ldots, x_i, \ldots, x_n$, а в выборке N_2 он выступает в качестве случайной величины Y, принимающей значения $y_1, y_2, \ldots, y_i, \ldots, y_n$.

Поскольку случайные величины X и Y и их значения $(x_1, x_2, \ldots, x_i, \ldots, x_n)$, $(y_1, y_2, \ldots, y_i, \ldots, y_n)$ представляют один и тот же лингвистический признак L, для проверки той или иной гипотезы можно использовать отношения упорядоченности типа $x_i < y_i$, $x_i > y_i$. На анализе этих отношений строятся два порядковых критерия — критерий знаков и критерий Вилкоксона.

2. Критерий знаков. Подсчет появлений лингвистического признака A в каждой из n порций выборок N_1 и N_2 дает два ряда независимых частот:

$$x_1, x_2, \ldots, x_i, \ldots, x_n; y_1, y_2, \ldots, y_i, \ldots, y_n,$$

где частоты, принадлежащие первой и второй выборкам, образуют пары (x_1, y_1), (x_2, y_2), ..., (x_i, y_i), ..., (x_n, y_n).

Далее составляются разности

$$z_1 = x_1 - y_1, z_2 = x_2 - y_2, \ldots, z_i = x_i - y_i, \ldots, z_n = x_n - y_n,$$

которые являются случайными величинами.

Исключив из анализа возможность равенства $x_i = y_i$ (т. е. $z_i = 0$; обоснование этого приема см. в [61, с. 321 и сл.]), будем рассматривать соотношение числа положительных значений z_i (суммы плюсов), равного m, с количеством отрицательных значений z_i (числом минусов), равным $n - m$. Величина m рассматривается в качестве численного значения критерия знаков.

Сформулируем сначала нулевую гипотезу H_0, которая утверждает, что в каждой i-й паре порций величины X и Y являются независимыми и одинаково распределенными, а вероятность того, что разность $z_i = x_i - y_i$ будет положительной, равна вероятности того, что эта разность окажется отрицательной, т. е.

$$P(z_i > 0) = P(z_i < 0) = 1/2,$$

с учетом того, что $P(z_i = 0) = 0$. Напротив, согласно альтернативной гипотезе H_1 различия между x_i и y_i являются значимыми, или иными словами,

$$P(z_i > 0) \neq P(z_i < 0).$$

Чтобы решить, какую из двух гипотез следует принять, необходимо опытное значение критерия знаков сопоставить с границами доверительного интервала для числа плюсов. Эти границы, характеризуемые уровнем существенности (уровнем значимости) $\alpha = q$, определяются исходя из следующих соображений.

Вероятность того, что n разностей z_i дадут m плюсов, описывается биномиальным распределением вида

$$P_n(m) = C_n^m \left(\frac{1}{2}\right)^m \left(1 - \frac{1}{2}\right)^{n-m} = \frac{n!}{m!\,(n-m)!} \cdot \frac{1}{2^n}. \qquad (9.3)$$

Дальнейшая процедура проверки гипотезы H_0 зависит от того каким уровнем значимости — односторонним $(q/2)$ или двусторонним (q) — мы будем пользоваться.

Рассмотрим сначала односторонний уровень значимости. Если справедливо равенство (9.3), то вероятность события, состоящего в том, что среди всех $z_1, z_2, \ldots, z_n$ количество положительных z_i, равное m, окажется больше некоторого граничного числа $\overline{m}$, составит

$$P_n(m \geqslant \overline{m}) = \sum_{m=\overline{m}}^{n} P_n(m) = \sum_{m=\overline{m}}^{n} C_n^m \frac{1}{2^n} \leqslant \frac{q}{2}. \qquad (9.4)$$

Одновременно вероятность события, заключающегося в том, что среди всех $z_1, z_2, \ldots, z_n$ число положительных z_i, равное m, будет меньше граничного $\underline{m}$, есть

$$P_n(m \leqslant \underline{m}) = \sum_{m=0}^{\underline{m}} P_n(m) = \sum_{m=0}^{\underline{m}} C_n^m \frac{1}{2^n} \leqslant \frac{q}{2}. \qquad (9.5)$$

Возьмем теперь величину $\overline{m}$ в качестве наименьшего количества положительных z_i, т. е. числа, для которого равенство (9.4) еще не превосходит некоторой вероятности $q/2$. Тогда гипотезу H_0 мы будем отвергать в тех случаях, когда число положительных z_i окажется больше, чем $\overline{m}$. При этом вероятность отвергнуть гипотезу H_0, когда она правильна, не превзойдет вероятности $q/2$.

Мы будем отвергать гипотезу H_0 и тогда, когда $m < \underline{m}$. И в этом случае вероятность отвергнуть гипотезу H_0, когда она правильна, не превышает вероятности $q/2$.

Оба эти правила, взятые порознь, представляют собой *односторонний критерий знаков*.

При *двустороннем критерии* задаются как верхняя, так и нижняя границы доверительного интервала. При этом гипотеза H_0 отвергается не только тогда, когда m — количество положительных z_i превышает границу $\overline{m}$, но также и тогда, когда число отрицательных z_j, равное $n - m$, оказывается ниже границы $\underline{m}$. Если границы $\overline{m}$ и $\underline{m}$ остались теми же, что и при одностороннем критерии, то уровень значимости двустороннего критерия равен $2 \cdot q/2 = q$. Значения границ для одностороннего и двустороннего критерия знаков см. в табл. VIII на стр. 370—372.

Теперь, пользуясь описанным математическим аппаратом, проверим лингвистическую гипотезу, согласно которой две достаточно большие, взятые наугад из одной и той же разновидности языка текстовые выборки окажутся идентичными относительно употребляемой данной разновидности лексики. Если проверка этой гипотезы с помощью критерия знаков дает положительный результат, то это позволит предполагать, что в указанных выборках употребление лексики подчиняется некоторой вероятностной норме.

Для проверки указанной гипотезы используем статистику 600 словоформ в двух случайно выбранных английских газетных текстах, по 100 тыс. словоупотреблений каждый.

В качестве образца осуществим полную проверку гипотезы H_0 об идентичности этих выборок относительно словоформы government 'правительство'.

Таблица 9.2

Проверка гипотезы об идентичности двух выборок английского публицистического текста относительно существительного government 'правительство'

Порции I и II выборок	1	2	3	4	5	6	7	8	9	10
Частоты $F_i^{\mathrm{I}} = x_i$ в I выборке	1	3	12	2	14	3	12	5	8	4
Частоты $F_i^{\mathrm{II}} = y_i$ во II выборке	2	4	1	7	1	6	3	10	5	8
Разности $z_i = x_i - y_i$	—1	—1	11	—5	13	—3	9	—5	3	—4

Продолжение табл. 9.2

Порции I и II выборок	11	12	13	14	15	16	17	18	19	20
Частоты $F_i^{\mathrm{I}} = x_i$ в I выборке	7	3	6	10	5	3	4	9	4	8
Частоты $F_i^{\mathrm{II}} = y_i$ во II выборке	14	1	4	10	13	7	8	6	7	3
Разности $z_i = x_i - y_i$	—7	2	2	0	—8	—4	—4	3	—3	5

Каждая из рассматриваемых выборок разделена на 20 порций, в которых определены частоты употребления government (см. вторую и третью строки табл. 9.2); в четвертой строке помещены разности $z_i = F_i^{\mathrm{I}} - F_i^{\mathrm{II}}$. Всего имеем 10 плюсов, 9 минусов, а в 14-й порции $z_{14} = 0$; исключив последний случай из рассмотрения, получаем $m = 10$, $n = 19$.

С помощью табл. VIII убедимся, что при уровне значимости $q = 0{,}05$ полученное нами из опыта значение критерия знаков $m = 10$ попадает внутрь доверительного интервала ($\underline{m} = 5$, $\overline{m} = 14$). Это говорит о том, что расхождения в частотах government по обеим выборкам несущественны, т. е. рассматриваемая словоформа имеет в обеих выборках постоянную вероятность. Аналогичным образом

исследовано употребление остальных 588 именных, глагольных, адвербиальных и служебных словоформ в английских публицистических текстах [32 а, с. 28]. Выяснилось, что только 4 словоформы — busy 'сыщик', refused 'отказал' can 'могут', by 'при, около' дают для критерия знаков такие значения, которые выходят за пределы 5%-ного доверительного интервала. Иными словами, только 0,7% обследованных словоформ обнаруживают неустойчивость своих вероятностей: остальные 99,3% словоформ имеют постоянные вероятности, что является косвенным указанием нормативности их употребления.

3. *U*-критерий Вилкоксона. Подобно критерию знаков, *критерий Вилкоксона* используется для проверки гипотез о несущественности расхождения двух лингвистических выборок, а в нашем случае для проверки гипотезы о нормативности. Строится этот критерий следующим образом. Пусть независимые выборки N_1 и N_2 разбиты соответственно на n_1 и на n_2 порций, причем интересующий нас лингвистический элемент встретился в i-й порции первой выборки x_i раз, а в i-й порции выборки N_2 он встретился y_i раз. Расположим теперь все значения x_i и y_i в одну строку в порядке возрастания численных значений x и y, не обращая внимания на индексы. В результате мы получаем смешанный вариационный ряд. *Инверсией* называется случай, когда y располагается перед x независимо от точного положения x_i и y_i в вариационном ряде.

Так, например, последовательность

$$y\ y\ x\ y\ x\ y\ x\ x$$

содержит 13 инверсий, поскольку первый x дает 2, второй 3, третий и четвертый по 4 инверсии. Полная сумма U числа инверсий в вариационном ряде есть случайная величина, численное значение которой, установленное в результате опыта, и является критерием Вилкоксона. Теперь попытаемся выяснить, о чем говорит численное значение U: когда оно требует принять, а когда отвергнуть проверяемую нулевую гипотезу?

Согласно критерию Вилкоксона, нулевая гипотеза должна быть отвергнута, если количество инверсий U выходит за некоторые пределы $\underline{u}$ и $\overline{u}$ (где $\underline{u} < \overline{u}$), зависящие от уровня значимости q. При этом оказывается, что если нулевая гипотеза верна, то случайная величина U имеет, как показал Б. Л. ван дер Варден [61, с. 336], распределение вероятностей с математическим ожиданием

$$M(U) = \frac{n_1 n_2}{2}$$

и дисперсией

$$D(U) = \sigma^2(U) = \frac{n_1 n_2}{12}(n_1 + n_2 + 1).$$

Показано [61, с. 337], что если $n_1 > 3$ и $n_1 + n_2 > 20$, то это распределение с достаточной точностью может считаться нормальным. Последнее обстоятельство позволяет определить пределы $\underline{u}$, $\overline{u}$,

если уровень значимости q задан. Введем нормированное уклонение

$$Z = \frac{U - M(U)}{\sigma(U)},$$

тогда

$$\underline{z} = \frac{\underline{u} - M(U)}{\sigma(U)}, \quad \bar{z} = \frac{\bar{u} - M(U)}{\sigma(U)}.$$

В силу вышесказанного случайная величина Z имеет нормальное распределение с параметрами: математическое ожидание 0 и дисперсия 1.

Уровень значимости q есть не что иное, как вероятность того, что вследствие случайных колебаний величина U выйдет за пределы $\underline{u}$ и $\bar{u}$, а Z — за пределы $\underline{z}$ и $\bar{z}$. Обычно выбирают $\underline{z} = -\bar{z}$. Тогда связь между $\underline{u}$, $\bar{u}$ и q может быть найдена из очевидных соотношений:

$$q = 1 - \frac{2}{\sqrt{2\pi}} \int_0^{\bar{z}} e^{-z^2/2}\, dz = 1 - 2\Phi(\bar{z}), \qquad (9.6)$$

$$\Phi(\bar{z}) = \frac{1-q}{2}, \qquad (9.7)$$

где $\Phi(\bar{z})$ — известный интеграл вероятностей. Таким образом, при заданном q соотношение (9.7) определяет $\bar{z}$ (см. табл. III на стр. 365 и табл. VI на стр. 369).

По значениям n_1, n_2, $\bar{z}$ можно легко определить пределы $\underline{u}$ и $\bar{u}$:

$$\underline{u} = -\bar{z}\sigma(U) + M(U) = -\bar{z}\sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)} + \frac{n_1 n_2}{2}, \qquad (9.8)$$

$$\bar{u} = \bar{z}\sigma(U) + M(U) = \bar{z}\sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)} + \frac{n_1 n_2}{2}. \qquad (9.9)$$

Если выбрать уровень значимости $q = 0{,}05$ (двустороннее ограничение), то это означает, что при справедливости нулевой гипотезы из 100 значений критерия Вилкоксона в среднем лишь пять могут выходить за пределы $\underline{u}$ и $\bar{u}$. Если же $q = 0{,}01$, то за пределы $\underline{u}$ и $\bar{u}$ может выходить лишь одно значение.

С помощью табл. VI мы убеждаемся, что если $q = 0{,}05$, то $\bar{z} = 1{,}96$, а если $q = 0{,}01$, то $\bar{z} = 2{,}58$. Эти значения $\bar{z}$ говорят о том, что выбранные нами уровни значимости хорошо согласуются с известным правилом «трех сигм», которое утверждает, что если некоторая случайная величина уклоняется на опыте от своего математического ожидания на величину, превышающую 3σ (при этом как раз $\bar{z} \geqslant 3$), то это происходит, как правило, за счет неслучайного воздействия на нее каких-то существенных факторов или за счет изменения условий наблюдения этой величины. Последнее принципиально меняет характер распределения вероятностей случайной величины.

С помощью критерия Вилкоксона можно проверить гипотезу об устойчивости вероятности словоформы government в английских публицистических текстах.

Снова возьмем две выборки N_1 и N_2 из английских газетных текстов. Каждую выборку разобьем на 20 порций, а в каждой порции определим частоту употребления контрольной словоформы (см. табл. 9.2).

Расположим теперь все значения $F_i^{\mathrm{I}} = x_i$ и $F_i^{\mathrm{II}} = y_i$ в порядке возрастания численных значений F, не обращая внимания на верхние и нижние индексы. В тех случаях, когда величины $F_i^{\mathrm{I}} = F_i^{\mathrm{II}}$, вопрос об их взаимном расположении решается путем жеребьевки. Исходя из этих условий, получаем следующий вариационный вид:

$$F_3^{\mathrm{II}},\ F_1^{\mathrm{I}},\ F_5^{\mathrm{II}},\ F_{12}^{\mathrm{II}},\ F_4^{\mathrm{I}},\ F_1^{\mathrm{II}},\ F_6^{\mathrm{I}},\ F_{16}^{\mathrm{I}},\ F_7^{\mathrm{II}},\ F_2^{\mathrm{I}},\ F_{12}^{\mathrm{I}},\ F_{20}^{\mathrm{II}},$$

$$F_{10}^{\mathrm{I}};\ F_{19}^{\mathrm{I}},\ F_2^{\mathrm{II}},\ F_{17}^{\mathrm{I}},\ F_{13}^{\mathrm{II}},\ F_8^{\mathrm{I}},\ F_9^{\mathrm{II}},\ F_{15}^{\mathrm{I}},\ F_{18}^{\mathrm{I}},\ F_6^{\mathrm{II}},\ F_{18}^{\mathrm{II}},\ F_4^{\mathrm{II}},\ F_{11}^{\mathrm{I}},$$

$$F_{16}^{\mathrm{II}},\ F_{19}^{\mathrm{II}},\ F_{20}^{\mathrm{I}},\ F_{10}^{\mathrm{II}},\ F_{17}^{\mathrm{II}},\ F_9^{\mathrm{I}},\ F_{18}^{\mathrm{I}},\ F_{14}^{\mathrm{I}},\ F_8^{\mathrm{II}},\ F_{14}^{\mathrm{II}},\ F_3^{\mathrm{I}},\ F_7^{\mathrm{I}},\ F_{15}^{\mathrm{II}},\ F_{11}^{\mathrm{II}},\ F_5^{\mathrm{I}}.$$

В обозначениях x и y этот вариационный ряд выглядит так:

$$y,\ x,\ y,\ y,\ x,\ y,\ x,\ x\ y,\ x,\ x,\ y,\ x,\ x,\ y,\ x,\ y,\ x,\ y,\ x,\ x,\ y,\ y,\ y,\ x,$$
$$y,\ y,\ x,\ y,\ y,\ x,\ x,\ x,\ y,\ y,\ x,\ x,\ y,\ y,\ x.$$

По описанной выше методике подсчитаем число инверсий:

$$u = 1 + 3 + 4 + 4 + 5 + 5 + 6 + 6 + 7 + 8 + 9 + 9 + 12 + 14 + 16 + 16 + 16 + 18 + 18 + 20 = 197.$$

В нашем эксперименте $n_1 = n_2 = 20$, поэтому условия $n_1 > 3$ и $n_1 + n_2 \geqslant 20$ выполнены; найдем математическое ожидание и дисперсию:

$$M(U) = \frac{n_1 n_2}{2} = \frac{20 \cdot 20}{2} = 200;$$

$$\sigma(U) = \sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)} = \sqrt{\frac{20 \cdot 20}{12} 41} = 37.$$

В соответствии с выражениями (9.8) и (9.9) 5%-ные доверительные пределы для критерия Вилкоксона в нашем случае таковы:

$$\underline{u} = 200 - 1{,}96 \cdot 37 \approx 127; \quad \bar{u} = 200 + 1{,}96 \cdot 37 \approx 273.$$

Нетрудно заметить, что число инверсий $u = 197$ контрольного слова government попадает в только что указанный доверительный интервал. Это дает право с уверенностью в 95% утверждать, что расхождения между распределениями в обеих выборках имеют случайный, нелингвистический характер, а само употребление этой словоформы подчинено некоторой норме.

Исследование с помощью критерия Вилкоксона остальных 599 словоформ [32а, с. 16 — 26] показало, что только 16 словоформ (began, city, continue, economy, greater, in, information, mother, movement, outside, plans, real, that, through, whose, yet) дают значения Вилкоксона, выходящие за 5%-ные доверительные пределы.

Между тем нулевая гипотеза оставалась бы справедливой, если бы таких случаев было тридцать. На основании этого можно сделать вывод, что словоформы в английских публицистических текстах имеют устойчивые распределения $P(F)$. Если в этих текстах и встречаются лингвистические единицы с неустойчивыми распределениями, то вклад таких единиц не превышает $\frac{16}{600} \cdot 100\% = 2{,}7\%$.

§ 3. Проверка гипотез о характере расхождений статистических характеристик языков, функциональных стилей и подъязыков с помощью параметрических критериев

При рассмотрении лексикологических, фонологических и грамматических проблем, связанных с сопоставлением различных языков, подъязыков и функциональных стилей, приходится сравнивать абсолютные и относительные частоты употребления лингвистических единиц в разных стилях, подъязыках, художественных произведениях. Все эти оценки и сопоставления осуществляются путем проверки гипотезы о существенности расхождения между соответствующими параметрами распределений интересующей исследователя лингвистической единицы.

1. Может ли средняя длина словоформы быть статистической характеристикой стиля и языка? Пусть имеется генеральная лингвистическая совокупность, элементы которой распределены нормально и характеризуются математическим ожиданием $M(X) = \mu$. На основе выборочных наблюдений внутри этой совокупности получено значение средней арифметической $\bar{x}$, не совпадающее точно со значением μ. Необходимо решить вопрос о существенности расхождения величин $\bar{x}$ и μ.

Решение этой задачи сводится к проверке нулевой гипотезы H_0, состоящей в допущении, что расхождения между $\bar{x}$ и μ несущественны, т. е. что $\bar{x} = \mu$. В этом случае альтернативная гипотеза H_1 заключается в утверждении, что $\bar{x} \neq \mu$.

Выбор критерия и статистической характеристики для проверки гипотез зависит от того, известно или неизвестно нам среднее квадратическое отклонение σ.

Чаще всего величина σ остается неизвестной. В этом случае в качестве статистической характеристики выбирается не случайная величина $\bar{x}$, а величина

$$t = \frac{\bar{x} - \mu}{\hat{s}} \sqrt{N}, \qquad (9.10)$$

имеющая распределение Стьюдента с $\nu = N - 1$ степенями свободы (см. гл. 8, § 2, п. 3 и § 3, п. 1).

Исходя из рассуждений § 1, п. 2, имеем критическую область значений характеристики t с нижней границей $t_{q,\nu}$, где $q = \alpha$ — заданный уровень значимости*. Тогда областью приемлемости гипотезы H_0 служат абсолютные значения $|t| < t_{q;\,\nu}$, а вероятность принятия этой гипотезы равна

$$P\left(\left|\frac{\bar{x} - \mu}{\hat{s}} \sqrt{N}\right| < t_{q;\,\nu}\right) = 1 - \alpha = 1 - q.$$

Вероятность же отвержения гипотезы H_0 и принятия альтернативной гипотезы H_1 в этом случае составляет

$$P\left(\left|\frac{\bar{x} - \mu}{\hat{s}} \sqrt{N}\right| \geqslant t_{q;\,\nu}\right) = q$$

при двусторонней критической области и

$$P\left(\frac{\bar{x} - \mu}{\hat{s}} \leqslant -t_{q\,;\,\nu}\right) = P\left(\frac{\bar{x} - \mu}{\hat{s}} \geqslant t_{q;\,\nu}\right) = \frac{q}{2}$$

для каждой из односторонних критических областей.

Проверка гипотезы H_0 о несущественности расхождения величин $\bar{x}$ и μ сводится к следующему:

а) определению по заданному уровню значимости и количеству степеней свободы $\nu = N - 1$ величины $t_{q;\,\nu}$;

б) вычислению по формуле (9.10) статистической характеристики t;

в) сравнению величин t и $t_{q;\,\nu}$.

Сам же критерий принятия или отвержения нулевой гипотезы формулируется таким образом:

а) при $|t| < t_{q;\,\nu}$ гипотеза H_0 принимается как правдоподобная, при этом утверждается, что опытная средняя $\bar{x}$ и математическое ожидание $\mu = M(\bar{x})$ статистически неразличимы (т. е. если различия между ними и наблюдаются, то они вызваны несущественными причинами);

б) при $|t| \geqslant t_{q;\nu}$ гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1, утверждающая, что расхождения между $\bar{x}$ и $\mu = M(\bar{x})$ не могут рассматриваться как незначительные статистические флуктуации, а вызваны существенными лингвистическими причинами.

Описанная процедура проверки нулевой гипотезы носит название *t*-критерия, или **критерия Стьюдента.**

---

* Поскольку проверка статистических гипотез основывается на выборе критической области, определяемой уровнем существенности, при рассмотрении наших гипотез мы будем исходить не из надежности p, а из уровня значимости q, учитывая при этом, что $z_q = z_p$ и $t_{q;\,\nu} = t_{p;\,\nu}$.

Если для генеральной лингвистической совокупности известно не только математическое ожидание $M(\bar{x}) = \mu$ интересующего нас элемента, но и среднее квадратическое отклонение σ, то в качестве статистической характеристики следует брать не t, а значение

$$z = \frac{\bar{x} - \mu}{\sigma} \sqrt{N}. \qquad (9.11)$$

Дальнейший ход решения задачи точно совпадает с операциями оценки нулевой гипотезы с помощью t-критерия: сначала по заданному уровню значимости q с помощью табл. VI на стр. 369 определяют значение z_q, служащее нижней границей критической области, после чего находят значение z. Если $|z| \geqslant z_q$, то гипотезу H_0 о несущественности расхождений $\bar{x}$ и μ следует отвергнуть и принять альтернативную гипотезу H_1; если же $|z| < z_q$, то нулевая гипотеза H_0 принимается и разность $|\bar{x} - \mu|$ рассматривается как случайная статистическая флуктуация.

Описанная процедура называется Z-**критерием** проверки статистических гипотез, или **критерием нормального закона.**

В гл. 7 (см. § 4, п. 4) была выдвинута гипотеза о нормальности распределения средних длин словоформ в языках мира. Если эта гипотеза подтвердится (ср. п. 2, § 1), то можно считать, что расхождения между средними длинами словоформ по отдельным языкам несущественны.

Но значит ли это, что такая несущественность расхождения между языками мира будет наблюдаться и при сравнении средних длин словоформ одного функционального стиля?

Чтобы ответить на этот вопрос, рассмотрим расхождения между средней длиной словоформы в научно-технической и деловой речи шести славянских языков, равной 6,13 буквы при стандарте 0,43, с одной стороны, и средней длиной словоформы в указанных стилях всех языков мира, которая равна примерно 7 буквам, — с другой [7, с. 204].

Допустим, что средние длины словоформ в языках мира по интересующим нас стилям распределены нормально, причем $M(\bar{x}) = \mu = 7$, одновременно $\bar{x} = 6{,}13$, а $\hat{s} = 0{,}43$. Пусть гипотеза H_0 состоит в том, что средняя длина славянской словоформы практически соответствует средней длине словоформы в языках мира, т. е. $\bar{x}_{сл} = \mu$.

Проверим нашу гипотезу с помощью t-критерия. Для этого, учитывая, что $\nu = N - 1 = 6 - 1 = 5$, воспользуемся табл. VI, по которой найдем нижнюю границу критической области $t_{0,05;\,5} = 2{,}57$. Затем определим значение t:

$$t = \frac{\bar{x}_{сл} - \mu}{s} \sqrt{N} = \frac{6{,}13 - 7{,}0}{0{,}43} \sqrt{6} = -\frac{0{,}87}{0{,}43} \cdot 2{,}45 = -4{,}95.$$

Неравенство $|t| > t_{0,05;5}$ свидетельствует о том, что значение нашей статистической характеристики t, с помощью которой мы проверяем гипотезу о несущественности расхождений средних длин, попадает в критическую область, в связи с чем эта гипотеза должна быть отвергнута.

Проверим эту гипотезу с помощью Z-критерия, полагая при этом, что $\hat{s} = \sigma = 0{,}43$.

Для этого с помощью табл. VI находим нижнюю границу критической области, равную $z_{0,05} = 1{,}96$, с которой сопоставляем абсолютное значение z, вычисленное с помощью соотношения (9.11):

$$\frac{6{,}13 - 7}{0{,}43} \sqrt{6} = -\frac{0{,}87}{0{,}43} \cdot 2{,}45 = -4{,}95.$$

Здесь снова имеем неравенство $|z| > z_{0,05}$, свидетельствующее о том, что значение статистической характеристики z опять попадает в критическую область.

Таким образом, и в том и в другом случае гипотеза H_0 о несущественности различия между средней длиной словоформы в славянских деловых и научно-технических текстах и средней длиной словоформы в аналогичных текстах языков мира должна быть отвергнута. Более правдоподобной оказывается альтернативная гипотеза H_1, согласно которой расхождение между $\bar{x}_{сл}$ и $\mu = M(X)$ является существенным.

Это расхождение можно отнести, вероятно, за счет двух причин. Во-первых, славянские языки используют флективно-аналитическую технику оформления именных форм, составляющих значительную часть деловых и научно-технических текстов во всех языках. Такая техника не дает столь значительного удлинения основы, как агглютинация в тюркских, финноугорских и других языках, занимающих значительную статистическую долю во взятой нами выборке языков. Во-вторых, славянские языки в отличие от немецкого и некоторых финноугорских языков сравнительно мало пользуются словосложением при образовании научно-технических и административно-деловых терминов.

2. Существенны ли расхождения значений избыточности в разговорной, беллетристической и деловой речи? Избыточность является информационной характеристикой, в которой обобщаются различные статистико-дистрибутивные свойства текста (см. гл. 5, § 5, п. 6). Поэтому при сопоставительном изучении стилей или других разновидностей языка важно знать, относится ли наблюдаемое расхождение в значениях избыточности двух стилей к существенному или несущественному. Сопоставление этих значений можно осуществить, проверяя с помощью параметрических критериев гипотезы о несущественности расхождения двух средних. Рассмотрим математическую схему проверки такой гипотезы.

Пусть имеются две нормально распределенные случайные и независимые лингвистические выборки достаточно большого объема:

$x_1, x_2, \ldots, x_{N_1}$ и $y_1, y_2, \ldots, y_{N_2}$. Эти выборки взяты либо из одной, либо из разных генеральных совокупностей.

Согласно теореме Ляпунова (см. гл. 6, § 4, п. 4), средняя $\bar{x}$ первой выборки есть нормально распределенная случайная величина с параметрами $M(\bar{x}) = \mu_1$ и $D(\bar{x}) = \sigma_x^2/N_1$, а средняя $\bar{y}$ второй выборки аналогично является нормально распределенной случайной величиной с параметрами $M(\bar{y}) = \mu_2$ и $D(\bar{y}) = \sigma_y^2/N_2$.

Необходимо определить существенность расхождения средних $\bar{x}$ и $\bar{y}$. Для этого зададим нулевую гипотезу H_0, состоящую в предположении, что $\bar{x}$ и $\bar{y}$ различаются несущественным образом, т. е. их математические ожидания равны: $M(\bar{x}) = M(\bar{y})$, или $\mu_1 = \mu_2$. Принятие гипотезы H_0 означает, что выборки N_1 и N_2 принадлежат не к разным, а к одной и той же генеральной совокупности. Отклонение этой гипотезы свидетельствует о том, что расхождение между $\bar{x}$ и $\bar{y}$ существенно, $M(\bar{x}) \neq M(\bar{y})$, а выборки N_1 и N_2 взяты из разных генеральных совокупностей.

Для проверки гипотезы H_0 введем величину $\delta = \bar{x} - \bar{y}$, которая также является случайной величиной с математическим ожиданием

$$M(\delta) = M(\bar{x} - \bar{y}) = M(\bar{x}) - M(\bar{y}) = \mu_1 - \mu_2$$

и дисперсией

$$D(\delta) = \sigma_\delta^2 = D(\bar{x}) + D(\bar{y}) = \sigma_x^2/N_1 + \sigma_y^2/N_2.$$

Рассмотрим нормированную случайную величину

$$\frac{\delta - M(\delta)}{\sqrt{D(\delta)}},$$

которая также распределена по нормальному закону с математическим ожиданием 0 и дисперсией 1. В силу соотношения (6.127) можно считать, что

$$P\left\{\frac{|(\bar{x}-\bar{y})-(\mu_1-\mu_2)|}{\sqrt{\sigma_x^2/N_1+\sigma_y^2/N_2}} < z_p\right\} = 2\Phi(z). \qquad (9.12)$$

Выбор статистического критерия снова зависит от того, известны ли нам дисперсии $D(\bar{x}) = \sigma_x^2$ и $D(\bar{y}) = \sigma_y^2$ или нет.

Пусть σ_x^2 и σ_y^2 известны. Тогда в качестве статистической характеристики можно взять величину

$$z = \frac{\bar{x}-\bar{y}}{\sigma_{\bar{x}-\bar{y}}},$$

которая получается из (9.12) в предположении, что $\mu_1 = \mu_2$ (гипотеза H_0). Затем назначается уровень значимости q, по которому из табл. VI берется величина z_q, выступающая в качестве границы критической области. Если имеет место неравенство $|z| < z_q$,

т. е. статистическая характеристика лежит в области приемлемости гипотезы H_0, то это означает, что мы можем принять с надежностью μ гипотезу о том, что различие между математическими ожиданиями средних $\bar{x}$ и $\bar{y}$ имеет характер случайной статистической флуктуации, откуда $M(\bar{x}) = M(\bar{y})$. Из равенства математических ожиданий следует, что выборки, по которым получены эти средние, взяты из одной генеральной совокупности.

В том случае, если $|z| \geqslant z_q$ и значение статистической характеристики попадает в критическую область, гипотеза H_0 должна быть отвергнута в пользу альтернативной гипотезы H_1, согласно которой различия между $\bar{x}$ и $\bar{y}$ существенны, а $M(\bar{x}) \neq M(\bar{y})$. Принятие альтернативной гипотезы означает, что средние взяты из выборок, не принадлежащих одной генеральной совокупности.

Обычно теоретические дисперсии σ_x^2 и σ_y^2 в распределениях лингвистических объектов остаются неизвестными, а вместо них используются выборочные дисперсии $\hat{s}_x^2$ и $\hat{s}_y^2$. Поэтому проверка лингво-статистических гипотез осуществляется не с помощью Z-критерия, а путем применения критерия Стьюдента. Переход к этому критерию осуществляется здесь исходя из следующих соображений.

Сперва предполагается, что $\sigma^2 = \sigma_x^2 + \sigma_y^2$. Тогда для разности $\bar{x} - \bar{y}$ имеем дисперсию

$$D(\bar{x}-\bar{y}) = \sigma^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right),$$

а также нормально распределенную нормированную случайную величину

$$\frac{(\bar{x}-\bar{y})-M(\bar{x}-\bar{y})}{\sqrt{D(\bar{x}-\bar{y})}} = \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sigma\sqrt{1/N_1+1/N_2}}. \qquad (9.13)$$

Среднее квадратическое отклонение σ можно заменить стандартом $s_{x,y}$, полученным из величины $s^2_{\bar{x},\bar{y}}$, являющейся средневзвешенной выборочных дисперсий

$$\hat{s}_x^2 = \frac{1}{N_1-1}\sum_{l=1}^{N_1}(x_l-\bar{x})^2$$

и

$$\hat{s}_y^2 = \frac{1}{N_2-1}\sum_{l=1}^{N_2}(y_l-\bar{y})^2.$$

Иными словами,

$$s_{\bar{x},\bar{y}} = \sqrt{\frac{\sum_{l=1}^{N_1}(x_l-\bar{x})^2+\sum_{l=1}^{N_2}(y_l-\bar{y})^2}{N_1+N_2-2}} = \sqrt{\frac{(N_1-1)\hat{s}_x^2+(N_2-1)\hat{s}_y^2}{N_1+N_2-2}},$$

Заменяя в равенстве (9.13) σ на $s_{\bar{x},\bar{y}}$, можно прийти к величине

$$t^* = \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{s_{\bar{x},\bar{y}}\sqrt{1/N_1+1/N_2}},$$

имеющей распределение Стьюдента с $\nu = N_1 + N_2 - 2$ степенями свободы. Проверка нулевой гипотезы H_0, заключающейся в предположении, что $M(\bar{x}) = M(\bar{y})$ (т. е. $\mu_1 = \mu_2$), предусматривает определение величины

$$t = \frac{\bar{x}-\bar{y}}{s_{\bar{x},\bar{y}}\sqrt{1/N_1+1/N_2}} \qquad (9.14)$$

и сравнение ее с табличным значением $t_{q;\nu}$. Если $|t| < t_{q;\nu}$, то гипотеза H_0 принимается и различия между $\bar{x}$ и $\bar{y}$ рассматриваются как несущественные, т. е. $M(\bar{x}) = M(\bar{y})$, а выборки N_1 и N_2 считаются принадлежащими одной генеральной совокупности. Если же $|t| \geqslant t_{q;\nu}$, то гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1, согласно которой различия между $\bar{x}$ и $\bar{y}$ считаются существенными, т. е. $M(\bar{x}) \neq M(\bar{y})$, а выборки N_1 и N_2 принадлежат к разным лингвистическим генеральным совокупностям. Критерий Стьюдента может быть использован как при малых, так и при больших значениях N_1 и N_2.

Познакомившись со схемой сравнения двух средних, перейдем к исследованию расхождений между средними значениями избыточности в разговорной, беллетристической и деловой речи шести европейских языков.

Таблица 9.3

Сравнение избыточностей стилей по шести европейским языкам

Разновидности языка	Избыточность $(\bar{R}+\bar{R})/2$ по языкам						R	$\hat{s}$	$s_{\bar{x},\bar{y}}$	
	Русский	Польский	Английский	Немецкий	Французский	Румынский				
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	
Разговорная речь (р)	0,777	0,813	0,753	0,792	0,757	0,801	0,782	0,024	0,0255	0,0266
Беллетристика (б)	0,812	0,791	0,818	0,745	0,773	0,788	0,788	0,027	0,0290	
Деловая речь (д)	0,868	0,866	0,875	0,835	0,872	0,802	0,853	0,029		

С этой целью возьмем средние арифметические нижних и верхних значений избыточности $\bar{R}_i = (\underline{R}_i + \bar{R}_i)/2$ [столбцы (2) — (7) табл. 9.3] и будем рассматривать их как случайные значения нормально распределенной величины избыточности по каждому из указанных выше стилей. Вычислив среднюю избыточность R и стандарт $\hat{s}$ для каждого стиля [столбцы (8) и (9)], попытаемся выяснить, насколько существенны расхождения между значениями средней избыточности по каждому из названных стилей. С этой целью будем проверять с помощью критерия Стьюдента нулевую гипотезу H_0, согласно которой расхождения в значениях R_i и R_j являются несущественными. Уровень значимости примем равным 0,05, число степеней свободы в нашем случае составляет $\nu = 6 + 6 - 2 = 10$.

При этих условиях граница критической области равна $t_{0,05;10} = 2,23$.

Попарное сравнение средних избыточностей стилей с помощью выражения (9.14) дает следующие результаты.

1) Для пары разговорная речь — беллетристический стиль имеем

$$t_{р,\,б} = \frac{R_р - R_б}{s_{р,\,б}\sqrt{1/N_1+1/N_2}} = \frac{0,782-0,788}{0,0255\sqrt{1/6+1/6}} = -\frac{0,006\cdot\sqrt{3}}{0,0255} = -\frac{0,014}{0,0255} \approx -0,41.$$

В связи с тем, что

$$|t_{р,\,б}| < t_{0,05;\,10} = 2,23,$$

нулевая гипотеза о несущественности различий средней избыточности разговорного и беллетристического стилей в шести европейских языках принимается как вполне правдоподобная.

2) Для пары разговорная речь — деловая речь аналогичным образом находим

$$t_{р,\,д} = \frac{0,782-0,853}{0,0266\sqrt{1/6+1/6}} = -\frac{0,071\cdot 1,732}{0,0266} \approx -4,59.$$

Полученное значение t здесь выше порога критической области, т. е.

$$|t_{р,\,б}| > t_{0,05;\,10} = 2,23.$$

Это говорит о существенности расхождений между разговорной и деловой речью. Иными словами, гипотезу о несущественности расхождений избыточности указанных разновидностей следует отвергнуть.

3) Пара беллетристика — деловая речь дает

$$t_{б,\,д} = \frac{0,788-0,853}{0,0290\sqrt{1/6+1/6}} = -\frac{0,065\cdot 1,732}{0,0290} \approx -4,02.$$

Здесь снова имеем неравенство

$$|t_{б,\,д}| > t_{0,05;\,10},$$

говорящее о том, что значение t попадает в критическую область отвержения нулевой гипотезы.

К аналогичным результатам можно прийти, используя вместо критерия Стьюдента Z-критерий (в этом случае предполагается, что $\sigma_{\bar{x}} = s_{\bar{x}}$, а $\sigma_{\bar{y}} = s_{\bar{y}}$). Итак, попарное сравнение средних избыточно-

стей по трем стилям шести европейских языков показывает, что различия величин R письменной фиксации разговорной речи и беллетристической речи несущественны*, в то время как расхождения в значениях R для деловой речи, с одной стороны, и разговорно-беллетристической разновидности в ее письменной форме, с другой, существенны. Причины этих существенных расхождений следует искать в статистико-дистрибутивных характеристиках деловой речи, отражающих ее качественные особенности. Среди этих особенностей, противопоставляющих ее разговорной и беллетристической речи, основное место, очевидно, занимает использование в деловой речи большого числа клише и штампов, нормированность синтаксиса, а также более или менее фиксированный выбор лексики. Все это значительно увеличивает избыточность деловой речи.

3. Оценка лексических расхождений между публицистикой ГДР и газетными текстами ФРГ. Использование статистических критериев дает возможность не только исследовать нормированность текста или сравнивать разные функциональные стили по таким усредненным характеристикам, как избыточность. Метод статистической проверки гипотез дает возможность выявлять и оценивать качественные расхождения в лексике, грамматике и фонологии двух разновидностей одного языка или двух близкородственных языков.

Оценка этих расхождений осуществляется путем сравнения частостей употребления лингвистических единиц $L_1, L_2, \ldots, L_n$ в двух выборках N_1 и N_2, каждая из которых представляет определенный язык или его разновидность.

Для каждой лингвистической единицы L вычисляются две относительные частоты: $f_1 = F_1/N_1$ для первой и $f_2 = F_2/N_2$ для второй выборки. Кроме того, предполагается, что в первом говорении (т. е. языке, варианте или диалекте) единица L имеет вероятность p_1, а во втором — вероятность p_2.

Затем выдвигаются две гипотезы: нулевая H_0, предполагающая, что $p_1 = p_2$, и альтернативная H_1, утверждающая, что $p_1 \neq p_2$. Так как функция распределения разности $f_1 - f_2$ имеет меньшие скачки и меньшую асимметрию, чем функции распределений f_1 и f_2, то можно утверждать, что разность $f_1 - f_2$ имеет нормальное распределение с математическим ожиданием $M(f_1 - f_2) = p_1 - p_2$ и дисперсией** $\sigma^2_{p_1-p_2} = \sigma^2_1 + \sigma^2_2$.

* Если бы в ходе эксперимента по определению энтропии и избыточности текста можно было бы учесть такие дополнительные средства устной речи как жест, мимика, интонация, то не исключено, что мы получили бы иные оценки ее избыточности. В этом случае сравнение величин $R_р$, $R_б$ и $R_д$ могло бы привести к другим результатам.

** Хотя точные значения дисперсий нам неизвестны, но при достаточно большом N можно заменить неизвестные величины σ^2_1, σ^2_2 их эмпирическими оценками

$$\hat{s}^2_1 = \frac{f_1(1-f_1)}{N_1 - 1}, \quad \hat{s}^2_2 = \frac{f_2(1-f_2)}{N_2 - 1}.$$

Дисперсия же $\sigma^2_{p_1-p_2}$ заменяется в этом случае оценкой $\hat{s}^2 = \hat{s}^2_1 + \hat{s}^2_2$.

Из рассуждений, приведенных в гл. 6, § 4, п. 4, следует, что нормированное отклонение

$$\frac{(f_1 - f_2) - (p_1 - p_2)}{\sigma_{p_1-p_2}}$$

распределено нормально. Отсюда при заданном уровне значимости q имеем

$$P\left\{\frac{|(f_1 - f_2) - (p_1 - p_2)|}{\sigma_p} < z_q\right\} = p.$$

Если величина z такова, что

$$|z| = \frac{|f_1 - f_2|}{\sigma_{p_1-p_2}} < z_q,$$

то можно считать, что вероятности p_1 и p_2 статистически неразличимы (т. е. $p_1 = p_2$). Если же $|z| > z_q$, то гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1, утверждающая, что $p_1 \neq p_2$.

С помощью описанного аппарата А. С. Ротарь [33, с. 163 — 199] исследовала соотношение частостей у 1000 наиболее частых словоформ в двух выборках немецких публицистических текстов. Первая выборка охватывает газетные тексты ГДР, а вторая — тексты из газет ФРГ за 1965 — 1969 г. Каждая выборка содержит 100 тыс. словоупотреблений. Для повышения надежности исследования взят трехсигмовый критерий ($z_q = 3$, $p = 0{,}9973$, $q = 0{,}0027$).

Всю процедуру определения существенности расхождения частостей в обоих литературных вариантах рассмотрим на примере словоформы nach 'после, через, по', которая, согласно данным частотного словаря А. С. Ротарь, имеет в выборке ГДР $F_1 = 317$ ($f_1 = 0{,}00317$), а в выборке ФРГ дает $F_2 = 538$ ($f_2 = 0{,}00538$).

1) После определения частостей вычисляются дисперсия и среднее квадратическое отклонение по каждой выборке в отдельности и по обеим выборкам, взятым вместе:

$$\sigma^2_1 = \hat{s}^2_1 = \frac{f_1(1-f_1)}{N_1} = \frac{0{,}00317 \cdot 0{,}99683}{10^5} = \frac{0{,}0032}{10^5},$$

$$\sigma^2_2 = \hat{s}^2_2 = \frac{0{,}00538 \cdot 0{,}99462}{10^5} = \frac{0{,}0053}{10^5};$$

$$\sigma^2_{p_1-p_2} = (0{,}0032 + 0{,}0053)/10^5 = 0{,}0085/10^5;$$

$$\sigma_{p_1-p_2} = \sqrt{0{,}0085/10^5} = 0{,}0003.$$

2) Вычисляется абсолютная величина z:

$$|z| = |0{,}00317 - 0{,}00538|/0{,}0003 = 7{,}03.$$

Поскольку $|z| > 3$, следует отвергнуть гипотезу о том, что предлог nach имеет одинаковую вероятность употребления в публицистических текстах ГДР и ФРГ. Более правдоподобной является гипотеза о существенном расхождении этих вероятностей, которое следует искать в дистрибутивно-статистических особенностях употребления этого предлога в восточно- и западногерманской публицистической прозе (ср. ниже).

По только что описанной схеме были сопоставлены частости остальных словоформ списка. Это сопоставление показало следующие результаты.

1. Более восьмисот словоформ дало либо полное совпадение частостей в обеих выборках, либо показало несущественные расхождения величин f_1 и f_2. В эту группу вошли такие общеупотребительные существительные, прилагательные и глаголы, как Land 'страна', Mittel 'средство', Ziel 'цель', kleinen 'малые', letzte 'последний', brauchen 'нуждаться'. Общенемецкие дистрибутивно-вероятностные нормы употребления показали также многие артиклевые формы, предлоги, союзы и вспомогательные формы глагола.

2. Существенное расхождение частостей f_1 и f_2 обнаружено почти у двухсот словоформ, покрывающих около 10% текста. Разные вероятности употребления в газетах ГДР и ФРГ имеют сложносокращенные словоформы и буквенные аббревиатуры (см. табл. 9.4), а также существительные и прилагательные терминологического значения, которые обозначают общественно-политические и экономические понятия, использующиеся преимущественно либо в ГДР, либо в ФРГ (табл. 9.5). Разные вероятности употребления дают географические названия (табл. 9.6), что объясняется либо расположением самих географических объектов, либо нормами географических терминологий, используемых в ГДР и ФРГ, либо преобладанием определенной тематики на страницах восточно- или западногерманской прессы.

Менее очевидны причины, порождающие существенные расхождения в частостях количественных числительных и названий месяцев (табл. 9.7), а также у глаголов (табл. 9.8), личных местоимений (табл. 9.9), предлогов и союзов (табл. 9.10). Эти вероятностные различия языка можно отнести за счет разной стилевой ориентации норм восточно- и западногерманских газет. В газетах ГДР печатаются тексты речей, произносимых от первого лица, здесь широко используются нормы деловой речи с ее именными конструкциями; даты, включая названия месяцев, чаще всего записываются в цифровой форме. Газеты ФРГ больше ориентируются на беллетристическую речь, поэтому в западногерманской публицистике преобладают глагольные обороты со сложным предложным управлением, а также широко используются сочинительные и подчинительные конструкции.

Таблица 9.4

№	Аббревиатуры и словоформы	f_1 (ГДР)	f_2 (ФРГ)	$\lvert z \rvert$
1	DDR 'ГДР'	0,00250	0,00005	15,3
2	DM 'марка ГДР'	0,00021	0,00089	6,5
3	USA 'США'	0,00093	0,00039	4,9
4	NATO 'НАТО'	0,00021	0,00002	3,8
5	Prof. 'проф.'	0,00027	0,00077	4,9
6	SED 'СЕПГ'	0,00029	0,00009	3,3
7	FDJ 'Союз немецкой молодежи'	0,00032	—	—

Таблица 9.5

№	Словоформы	f_1	f_2	$\lvert z \rvert$
1	Republik 'республика'	0,00113	0,00007	9,6
2	Genosse 'товарищ'	0,00047	0,00001	6,7
3	Frieden 'мир'	0,00055	0,00004	6,4
4	Einheit 'единство'	0,00038	0,00003	5,5
5	Socialismus 'социализм'	0,00038	0,00003	5,5
6	sozialistischen 'социалистические'	0,00055	0,00008	5,9
7	Betriebe 'предприятия'	0,00037	0,00003	5,4
8	demokratischen 'демократические'	0,00042	0,00003	5,8
9	Friedens 'мира'	0,00036	0,00002	5,6
10	Betrieb 'предприятие'	0,00043	0,00005	5,4
11	Werktätigen 'трудящиеся'	0,00034	0,00002	5,3
12	Zone 'зона'	0,00001	0,00030	5,3
13	Staaten 'государство'	0,00087	0,00030	5,2
14	Prozente 'проценты'	0,00034	0,00005	4,8
15	Steigerung 'повышение'	0,00027	0,00003	4,4
16	Stellvertreter 'заместитель'	0,00027	0,00003	4,4
17	Polizei 'полиция'	0,00005	0,00032	4,4
18	Bundesregierung 'правительство ФРГ'	0,00007	0,00038	4,4
19	Konferenz 'конференция'	0,00006	0,00032	4,3
20	Bau 'стройка'	0,00004	0,00025	4,0
21	Ministerrates 'Совета Министров'	0,00022	0,00001	4,2
22	Arbeiterklasse 'рабочий класс'	0,00027	—	—
23	Staatsrates 'Государственного Совета'	0,00022	—	—
24	Volkskammer 'Народная палата'	0,00022	—	—

Таблица 9.6

№	Словоформы	f_1	f_2	$\lvert z \rvert$
1	Bonner 'бонский'	0,00069	0,00009	6,8
2	Westberlin 'Западный Берлин'	0,00073	0,00013	6,6
3	Westdeutschland 'Западная Германия'	0,00060	0,00006	6,6
4	München 'Мюнхен'	0,00001	0,00039	6,3
5	Berlin 'Берлин'	0,00070	0,00152	5,5
6	westdeutschen 'западногерманские'	0,00054	0,00009	5,5
7	amerikanischen 'американские'	0,00016	0,00055	4,8
8	Bremen 'Бремен'	—	0,00028	
9	Istambul 'Стамбул'	—	0,00021	
10	Nill 'Нил'	—	0,00099	

Таблица 9.7

№	Словоформы	f_1	f_2	$\lvert z \rvert$
1	drei 'три'	0,00013	0,00093	8,1
2	zwei 'два'	0,00026	0,00108	7,1
3	elf 'одиннадцать'	0,00001	0,00032	5,7
4	sieben 'семь'	0,00002	0,00025	4,6
5	fünf 'пять'	0,00011	0,00042	4,4
6	sechs 'шесть'	0,00009	0,00039	4,3
7	vier 'четыре'	0,00012	0,00035	3,3
8	August 'август'	0,00011	0,00086	7,5
9	Februar 'февраль'	0,00008	0,00048	5,7
10	September 'сентябрь'	0,00008	0,00035	4,1
11	Juli 'июль'	0,00006	0,00023	3,4

Таблица 9.8

№	Словоформы	f_1	f_2	$\lvert z \rvert$
1	wird 'становится'	0,00356	0,00487	4,4
2	seien 'будьте'	0,00012	0,00042	4,3
3	sieht 'видит'	0,00009	0,00032	3,8
4	unternehmen 'предпринимать'	0,00001	0,00020	4,2
5	könnte 'мог бы'	0,00018	0,00045	3,4
6	dürfte 'мог бы'	0,00003	0,00019	3,2

Таблица 9.9

№	Словоформы	f_1	f_2	$\lvert z \rvert$
1	wir 'мы'	0,00454	0,00096	15,3
2	unserer 'наших'	0,00166	0,00013	11,7
3	ich 'я'	0,00271	0,00067	11,3
4	alle 'все'	0,00239	0,00065	10,0
5	unsere 'наши'	0,00143	0,00015	10,0
6	mir 'мне'	0,00045	0,00007	5,4
7	du 'ты'	0,00057	0,00009	6,0
8	unser 'наш'	0,00050	0,00019	3,9

Таблица 9.10

№	Словоформы	f_1	f_2	$\lvert z \rvert$
1	nach (prp) 'после, через, по, согласно'	0,00317	0,00538	7,6
2	bis 'пока'	0,00004	0,00046	6,0
3	zurück 'назад, обратно'	0,00011	0,00059	6,0
4	zur (zu+der) 'к, за, в'	0,00191	0,00312	5,4
5	da 'так как'	0,00009	0,00042	4,7
6	am (an+dem) 'у, на, к'	0,00277	0,00384	4,1
7	bei 'у, при, возле, около'	0,00017	0,00042	3,1

§ 4. Проверка статистических гипотез о тождестве двух лингвистических распределений

1. Сравнение эмпирического и теоретического или двух эмпирических распределений. При решении некоторых теоретических и прикладных вопросов, например при формальном выделении в тексте ключевых и терминологических слов, возникает необходимость рассмотреть не только параметры, но и характер всего лингвистического распределения. Эта задача также решается путем проверки статистических гипотез о тождестве двух эмпирических распределений (вариационных рядов) или об идентичности эмпирического и теоретического распределения.

Схема проверки тождества эмпирического и теоретического распределений выглядит следующим образом. Пусть имеется эмпирическое распределение, например распределение относительных (f), относительных накопленных (f^*), абсолютных (F), абсолютных накопленных (F^*) частот лингвистического признака L в выборочной совокупности текстов. Это распределение сопоставляется с соответствующим ему гипотетическим теоретическим распределением — с распределением вероятностей (p), накопленных вероятностей (p^*), математических ожиданий частот $[M\ (F)]$, математических ожиданий накопленных частот $[M\ (F^*)]$ признака L в генеральной совокупности текстов. Выдвигается нулевая гипотеза H_0, состоящая в том, что эмпирическое распределение выборки соответствует теоретическому распределению генеральной совокупности. Сам закон распределения может быть задан различным образом, например в форме плотности распределения $f\ (x)$, в интегральной форме распределения $F\ (x)$ или каким-то другим образом.

Для того чтобы принять или отвергнуть гипотезу H_0, конструируется некоторая статистическая характеристика W, определяющая степень расхождения эмпирического и теоретического распределений. Величина W может быть построена различными способами: в качестве ее рассматривается либо сумма квадратов отклонений

частот от их математических ожиданий $W = \Sigma\,[F - M\,(F)]^2$, либо сумма тех же квадратов с их весами $W' = \Sigma\, c\,[F - M(F)]^2$, либо максимальное отклонение интегральной эмпирической функции $F_N\,(x)$ от интегральной теоретической функции распределения $F\,(x)$ и т. д.

Сконструированная одним из перечисленных способов статистическая характеристика W является случайной величиной, закон распределения которой зависит от закона распределения случайной величины X ($X = F, f, F^*, f^*$) и от числа испытаний N. Если гипотеза H_0 верна, то закон распределения выборочной характеристики W определяется законом теоретического распределения случайной величины X и числом N.

Будем считать, что закон распределения известен. Пусть для данной выборки с числом опытов N мера расхождения W приняла значение w. Возникает вопрос: определяется ли то или иное численное значение величины W случайными несущественными флуктуациями (в этом случае гипотеза H_0 принимается) или это отклонение настолько значительно, что его следует относить за счет существенного расхождения между теоретическим и эмпирическим распределениями (в этом случае гипотеза H_0 отвергается)?

Чтобы ответить на этот вопрос, определим, предполагая, что гипотеза H_0 верна, вероятность того, что опытное значение w не превысит некоторого численного порога w^* выбранной меры расхождения — порога, выше которого расхождения между эмпирическим и теоретическим распределениями следует считать существенными.

Если вероятность $P\ (w < w^*)$ велика, то гипотезу H_0 можно принять как вполне правдоподобную. Если же указанная вероятность очень мала, то следует признать, что экспериментальные данные противоречат гипотезе H_0, и ее нужно отклонить. Аналогичным образом проверяется гипотеза о тождестве двух эмпирических распределений.

Как же следует строить статистическую характеристику W? Поскольку функции распределения $f\,(x)$, $F\,(x)$ чаще всего остаются неизвестными, целесообразно выбирать величину W таким образом, чтобы закон ее распределения не зависел бы от указанных функций распределений и их параметров. Иными словами, для проверки гипотез о расхождениях распределений целесообразно применять непараметрические критерии согласия, среди которых наиболее употребительным является критерий χ^2 Пирсона. Кроме того, здесь применяется критерий Колмогорова, упрощенные критерии Романовского и Ястремского, проверка нормальности распределения с помощью чисел Вестергарда и сетки Турбина, критерий Колмогорова — Смирнова.

2. Критерий χ^2 Пирсона. Пусть в результате N выборочных наблюдений исследуемая лингво-статистическая случайная величина X принимает определенные значения (варианты), которые сгруппированы в k интервалах статистического ряда, представленного в табл. 9.11.

Таблица 9.11

Интервалы вариант	(x_1, x_2)	(x_2, x_3)	...	(x_k, x_{k+1})	Σ
Частота (число значений X, попавших в интервал)	F_1	F_2	...	F_k	N
Частость $f_i = F/N$	f_1	f_2	...	f_k	1
Математическое ожидание частоты	Np_1	Np_2	...	Np_k	N
Вероятность попадания X в данный интервал	p_1	p_2	...	p_k	1

Предположим, что нам известен закон распределения случайной величины X, который задан интегральной функцией $F\,(x)$ или плотностью $f\,(x)$. Зная этот закон, можно определить вероятность p_i попадания случайной величины в каждый из интервалов (см. нижнюю строку табл. 9.11).

Задача состоит в том, чтобы проверить нулевую гипотезу H_0, утверждающую, что распределение полученных из опыта частостей f_i (или частот F_i) согласуется с теоретическим распределением вероятностей p_i (или соответственно математических ожиданий Np_i).

Будем проверять гипотезу H_0 о согласованности эмпирического и теоретического распределений, исходя из расхождений между частостями f_i и вероятностями p_i. Если эти расхождения малы, то здравый смысл подсказывает нам, что гипотезу H_0 можно считать правдоподобной. Если же расхождения велики, то ее вероятно, следует отвергнуть.

Мера расхождения эмпирического и теоретического распределений частостей и вероятностей, известная под названием **критерия** χ^2 (хи-квадрат) **Пирсона**, задается равенством

$$\chi^2 = \sum_{i=1}^{k} \frac{N}{p_i}(f_i - p_i)^2 = N \sum_{i=1}^{k} \frac{(f_i - p_i)^2}{p_i}, \qquad (9.15)$$

где частное N/p_i является весовым коэффициентом. Введение этого коэффициента объясняется тем, что отклонения, относящиеся к разным интервалам, нельзя считать равноправными: одна и та же абсолютная величина отклонения $f_i - p_i$ может быть очень незначительной при большом значении p_i и, наоборот, весьма заметной при очень малом p_i. Поэтому в качестве весовых коэффициентов берут величины, обратно пропорциональные вероятностям p_i.

Если оценивать соответствие эмпирического и теоретического распределений по расхождениям частот математических ожиданий $F_i - Np_i$, то легко показать, что выражение (9.15) примет вид

$$\chi^2 = \sum_{i=1}^{k} \frac{(F_i - Np_i)^2}{Np_i}. \qquad (9.16)$$

При построении интервальных вариационных рядов частоту лингвистического признака в данном интервале мы обозначали символом n_i, в связи с этим формула (9.16) может быть представлена также в виде

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i}. \qquad (9.17)$$

Случайная величина χ^2, представляющая собой сумму квадратов величин, связанных одной линейной зависимостью (напомним, что $\Sigma f_i = 1$, а $\Sigma F_i = N$) имеет важное свойство, заключающееся в том, что ее *распределение не зависит от закона распределения $F(x)$ или $f(x)$ и от N, а зависит от числа интервалов k*; в связи с этим при $N \to \infty$ распределение критерия χ^2 стремится к уже известному распределению χ^2 с $k - 1$ степенями свободы (ср. гл. 8, § 3, п. 1).

Сходимость распределения критерия χ^2 к χ^2-распределению дает возможность построить, задав определенный уровень значимости, критическую область для проверки гипотезы о соответствии эмпирического и теоретического распределений. Действительно, если величина χ^2 равна нулю, то это значит, что все квадраты разностей $(F_i - Np_i)^2$ или $(f_i - p_i)^2$ равны нулю, т. е. опытные частоты, точно соответствуя теоретическим, дают полное совпадение эмпирического и теоретического распределений. В остальных случаях величина χ^2 отлична от нуля и становится тем больше, чем больше растут расхождения между указанными частотами и их распределениями. Когда величина расхождения достигает определенного уровня, значение χ^2 переходит из области принятия гипотезы H_0 в критическую область ее отвержения.

Схема проверки нулевой гипотезы H_0 с помощью критерия χ^2 выглядит следующим образом.

1. На основании предшествующего опыта выбирается предполагаемый закон распределения изучаемой лингвистической величины X и определяются параметры этого закона.

2. Полученные из эксперимента частости (частоты) группируются в интервалы (группы), при этом малочисленные частоты объединяются в один интервал*.

* Математический смысл этого объединения состоит в следующем. При использовании критерия χ^2 следует иметь в виду, что биномиальное распределение частот $F_i = n_i$ сходится к нормальному. Этот предельный переход происходит достаточно быстро только тогда, когда и вероятность p, и вероятность q не слишком малы. Из этого следует, что математически корректным является такое применение критерия χ^2, при котором ни одна из интервальных частостей (соответственно вероятностей) и частот (соответственно математических ожиданий) не слишком мала. Это и заставляет нас объединять частоты крайних интервалов. Обычно объединяются интервалы, имеющие $F_i < 5$, с тем чтобы иметь в новых интервалах $F_i \geqslant 5$. Некоторые авторы требуют, чтобы теоретическая и эмпирическая частота интервала была не менее 10.

Существуют приемы, которые позволяют устранить неточности, возникающие в связи с применением непрерывного распределения χ^2 к дискретному распределению интервалов (*поправка на непрерывность Йетса*), а также коррективы, связанные с завышенным значением величины χ^2, вычисленной по сгруппированным данным (*поправки Шеппарда*). Подробнее об этом см. [7, с. 232]; [61, с. 284].

3. На основе выбранного закона распределения вычисляются вероятности (математические ожидания).

4. По формулам (9.15) или (9.16) вычисляется значение критерия χ^2.

5. Определяется число степеней свободы $\nu = k - l$, где k — число интервалов, а l — число налагаемых связей (ср. гл. 8, § 3, п. 1).

6. По заданному уровню значимости и числу степеней свободы с помощью табл. V на стр. 368 находят пороговое значение $\chi^2_{q;\nu}$, отделяющее область приемлемости гипотезы H_0 от критической области ее отклонения.

7. Полученное в п. 4 значение χ^2 сравнивается с пороговым значением $\chi^2_{q;\nu}$. Если имеет место неравенство $\chi^2 < \chi^2_{q;\nu}$, т. е. значение критерия Пирсона лежит в области приемлемости гипотезы, то гипотеза H_0 о согласованности эмпирического и теоретического распределений считается принятой, а это означает, что расхождения в этих распределениях рассматриваются как несущественные. Если же значение χ^2 попадает в критическую область, т. е. $\chi^2 \geqslant \chi^2_{q;\nu}$, то нулевая гипотеза отвергается и расхождения в распределениях рассматриваются как существенные.

Оценку гипотезы H_0 можно осуществлять не только путем сравнения величин χ^2 и $\chi^2_{q;\nu}$, но и через определение вероятности того, что некоторая случайная величина, имеющая χ^2-распределение, примет при ν степенях свободы значение, не меньшее, чем вычисленная по формуле (9.16) величина критерия χ^2. Если вероятность $P(\chi^2_\nu \geqslant \chi^2)$ окажется ниже некоторого заданного наперед уровня значимости, например ниже $q = 0{,}05$, то это означает, что вероятность случайных отклонений окажется очень малой величиной. Тогда мы должны признать нашу оценку χ^2 отклонения эмпирического распределения от теоретического закона неслучайной (ведь случайные явления с очень малой вероятностью следует считать практически невозможными). Это указывает на неправдоподобность гипотезы H_0 о несущественности расхождений между эмпирическим и теоретическим распределениями. Если же вероятность $P(\chi^2_\nu \geqslant \chi^2)$ достаточно велика (в нашем случае больше 0,05), то нулевую гипотезу о близости обоих распределений следует принять.

§ 5. Распределение средних длин словоформ в языках мира

1. Проверка гипотезы о нормальности распределения средних длин словоформ с помощью критерия χ^2 Пирсона. Чтобы осуществить проверку указанной гипотезы, необходимо сравнить полученные из опыта частости или частоты средних длин с их теоретическими вероятностями или математическими ожиданиями.

Вычисление этих теоретических величин может быть осуществлено двумя способами: исходя или из дифференциальной формы нормального закона (плотности вероятности), или из его интегральной формы.

Рассмотрим первый способ. Областью значений непрерывной случайной лингвистической величины X является либо вся числовая ось, либо часть ее. Разобьем эту область на m интервалов, соответствующих интервалам группировки выборочных измерений величины X. Каждый интервал ограничен нижней границей α_i и верхней границей β_i, причем $\beta_i = \alpha_{i+1}$.

Будем оперировать событиями $A_1, A_2, \ldots, A_m$, состоящими в том, что величина X попадает в интервалы с номерами 1, 2, ..., m. Вероятности появления указанных событий соответственно равны $p_1, p_2, \ldots, p_m$. Для выражения этих вероятностей используем значения плотности вероятности случайной величины. Используя рассуждения п. 3, 5 и 7 § 2 гл. 6, можно утверждать, что вероятность попадания случайной величины в интервал (α_i, β_i) равна произведению длины интервала $\beta_i - \alpha_i = \Delta x_i$ на значение плотности вероятности в одной из точек этого интервала. Такой точкой может быть середина интервала $(\beta_i - \alpha_i)/2 = \Delta x_i$.

Таким образом, имеем

$$p_i = P(\alpha_i < X < \beta_i) = \Delta x_i f(x) = (\beta_i - \alpha_i) f(x_i).$$

Наша случайная величина распределена по нормальному закону, поэтому

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

Отсюда следует, что

$$p_i = (\beta_i - \alpha_i) f(x_i) = \frac{\beta_i - \alpha_i}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/(2\sigma^2)}.$$

Поскольку значения μ, σ^2, σ обычно неизвестны, они заменяются величинами $\bar{x}$, $\hat{s}^2$, $\hat{s}$, полученными из эмпирического распределения. В результате приходим к выражению

$$p_i = \frac{\beta_i - \alpha_i}{\hat{s}\sqrt{2\pi}} e^{-(x_i-\bar{x})^2/(2\hat{s}^2)}. \qquad (9.18)$$

Учитывая, что $(x_i - \bar{x})/\hat{s} = z$ (см. гл. 6, § 4, п. 4), введем вспомогательную функцию

$$\varphi(z) = \frac{1}{2\pi} e^{-z^2/2},$$

применение которой дает возможность пользоваться при вычислениях табл. II на стр. 364.

Поскольку $\Delta x_i = \beta_i - \alpha_i$, перепишем (9.18) в виде

$$p_i = \frac{\Delta x_i}{\hat{s}} \varphi(z_i).$$

Теперь определим, чему равно математическое ожидание частоты $M(n_i) = n_i'$ появления величины X в i-м интервале.

Выбирая из непрерывной генеральной совокупности N наблюдений, мы формируем ситуацию, аналогичную той, при которой

признак в генеральной совокупности принимает конечное число значений и при которой имеют место m событий $A_1, A_2, \ldots, A_m$, образующих полную систему. Отсюда следует, что теоретическая частота (математическое ожидание) того, что при N испытаниях непрерывная случайная величина X находится в i-м интервале, исходя из (9.18), определяется равенством

$$n_i' = \frac{N\Delta x_i}{\hat{s}\sqrt{2\pi}} e^{-(x_i-\bar{x})^2/(2\hat{s}^2)}, \qquad (9.19)$$

где $p_i = n_i'/N$. Используя вспомогательную функцию $\varphi(z)$, перепишем выражение (9.19) в виде

$$n_i' = \frac{N\Delta x_i}{\hat{s}} \varphi(z_i). \qquad (9.20)$$

Обратимся теперь ко второму способу определения теоретических частот.

Исходя из свойства 4 интегральной функции распределения (см. гл. 6, § 2, п. 3), вероятность попадания случайной величины, подчиненной нормальному закону, на заданный интервал можно определить из выражения

$$p_i = P(\alpha_i < X < \beta_i) = F(\beta_i) - F(\alpha_i) = \Phi\left(\frac{\beta_i - \mu}{\sigma}\right) - \Phi\left(\frac{\alpha_i - \mu}{\sigma}\right).$$

Заменяя неизвестные μ и σ на опытные значения $\bar{x}$ и s (или $\hat{s}$), приходим к соотношению

$$p_i = \Phi\left(\frac{\beta_i - \bar{x}}{s}\right) - \Phi\left(\frac{\alpha_i - \bar{x}}{s}\right). \qquad (9.21)$$

Соответственно теоретическая частота того, что при N испытаниях случайная величина X попадет в i-й интервал, равна

$$n_i' = N\left[\Phi\left(\frac{\beta_i - \bar{x}}{s}\right) - \Phi\left(\frac{\alpha_i - \bar{x}}{s}\right)\right]. \qquad (9.22)$$

Применим описанную процедуру к проверке гипотезы о нормальности распределения средних длин словоупотреблений в языках мира (ср. гл. 7, § 4, п. 4). Эта проверка осуществляется по следующей схеме.

1) Значения средних длин словоупотреблений для 100 языков мира, использующих буквенную письменность (см. табл. 7.27 на стр. 259), сгруппированы в интервалы [столбец (1) табл. 9.12], каждый из которых имеет длину $\Delta x = 0{,}6$. Значения середин интервалов (x_i) указаны в столбце (2). Для каждого интервала указано число языков (n_i), средние длины словоупотреблений которых попадают в данный интервал [столбец (3)].

2) Среднюю арифметическую $\bar{x}$ и опытную дисперсию будем вычислять по методу моментов (см. гл. 7, § 3, п. 3), принимая за произвольное число a середину шестого интервала ($a = 5{,}70$).

Таблица 9.12

Номер интервала	$\alpha_i < X < \beta_i$	x_i	n_i	$x_i - a$	$\frac{x_i-a}{\Delta x}$	$n_i \frac{x_i-a}{\Delta x}$	$n_i \left(\frac{x_i-a}{\Delta x}\right)^2$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	2,40—3,00	2,70	0	—3,00	—5,0	0	0
2	3,00—3,60	3,30	2	—2,40	—4,0	— 8,0	32
3	3,60—4,20	3,90	8	—1,80	—3,0	—24,0	72
4	4,20—4,80	4,50	20	—1,20	—2,0	—40,0	80
5	4,80—5,40	5,10	21	—0,60	—1,0	—21,0	21
6	5,40—6,00	5,70	22	0	0	0	0
7	6,00—6,60	6,30	16	0,60	1,0	16,0	16
8	6,60—7,20	6,90	2	1,20	2,0	4,0	8
9	7,20—7,80	7,50	5	1,80	3,0	15,0	45
10	7,80—8,40	8,10	2	2,40	4,0	8,0	32
11	8,40—9,00	8,70	2	3,00	5,0	10,0	50
			100			—40,0	356

Подставляя в формулу (7.6) приведенные числовые значения и сумму из столбца (6), получаем значение средней арифметической:

$$\bar{x} = 5{,}70 + \frac{0{,}6}{100}(-40) = 5{,}46.$$

Учитывая тот факт, что все значения признака $\hat{x}_i$ имеют общий множитель Δx, воспользуемся для упрощения расчетов опытной дисперсии формулой (7.21). Подставив в нее уже приводившиеся выше числовые значения и сумму чисел столбца (7), имеем

$$s^2 = \frac{356}{100}\,0{,}36 - (5{,}46 - 5{,}70)^2 = 1{,}224.$$

Далее определяем несмещенную оценку выборочной дисперсии:

$$\hat{s}^2 = \frac{N}{N-1}\,s^2 = \frac{100}{99}\cdot 1{,}224 = 1{,}2364.$$

Следовательно, стандарт равен $\hat{s} = \sqrt{1{,}2364} \approx 1{,}11$.

3) Используя выражение (9.20), вычислим теоретические частоты n_i' для дифференциальной формы нормального закона. Расчеты показаны в столбцах (3) — (7) табл. 9.13, причем следует учитывать, что

$$\frac{\Delta x}{\sigma} = \frac{\Delta x}{\hat{s}} = \frac{0{,}60}{1{,}11} \approx 0{,}5405.$$

4) Получив значения n_i', можно непосредственно перейти к сравнению по критерию χ^2 эмпирических и теоретических частот. Результаты вычислительной работы, производимой по формуле (9.17), показаны в столбцах (9) — (11) табл. 9.13. Обращаем внимание читателя на укрупнение малочисленных интервалов [см. столбцы (2) и (7)].

Таблица 9.13

Номер интервала	$\hat{x}_i$	n_i	$x_i - \bar{x}$	$z_i = \frac{x_i-\bar{x}}{\hat{s}}$	$\varphi(z_i)$	$p_i = \frac{\Delta x}{\hat{s}}\varphi(z_i)$	$n_i' = Np_i$	n_i' (округл. до целых чисел)	$n_i - n_i'$	$(n_i - n_i')^2$	$\frac{(n_i-n_i')^2}{n_i}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	2,70	0	—2,76	—2,4865	0,0180	0,0097	0,97	1			
	3,30	2 (10)	—2,16	—1,9459	0,0608	0,0329	3,29 (12,35)	3	—2,35	5,5225	0,447
	3,90	8	—1,56	—1,4054	0,1497	0,0809	8,09	8			
2	4,50	20	—0,96	—0,8649	0,2756	0,1490	14,90	15	5,10	26,0100	1,742
3	5,10	21	—0,36	—0,3243	0,3790	0,2051	20,51	20	0,49	0,2401	0,012
4	5,70	22	0,24	0,2161	0,3894	0,2105	21,05	21	0,95	0,9025	0,043
5	6,30	16	0,84	0,7568	0,2989	0,1616	16,16	16	—0,16	0,0256	0,002
	6,90	2	1,44	1,2973	0,1714	0,0926	9,26	10			
	7,50	5 (11)	2,04	1,8378	0,0734	0,0397	3,97 (14,80)	4			
6	8,10	2	2,64	2,3784	0,0235	0,0127	1,27	1	3,80	14,4400	0,976
	8,70	2	3,24	2,9189	0,0056	0,0030	0,30	1			
		100				0,9974	$\sum_{i=1}^{6} p_i = 99{,}74$	100			3,222

В столбцах (4) — (6) можно брать также разность эмпирической частоты n_i и округленной до целых чисел теоретической частоты n_i' [см. столбец (8)].

5) Частоты всякого нормального распределения, устанавливаемого на основании наблюдаемого распределения, подчинены трем связям ($l = 3$). Во-первых, сумма наблюдаемых частот фиксирована, т. е. $\sum_{i=1}^{k} n_i = N$; во-вторых, теоретические частоты должны давать среднюю $\bar{x}$, равную средней эмпирических частот; в-третьих, дисперсии теоретического и эмпирического распределений должны быть также равны.

В результате укрупнения малочисленных интервалов общее их количество вместо одиннадцати становится равным шести ($k = 6$). В случае нормального распределения мы имеем дело с двумя параметрами $M(X) \approx \bar{x}$ и $D(X) = \sigma^2 \approx \hat{s}^2$, а также с суммой эмпирических частот $N = \sum_{i=1}^{k} n_i$. Таким образом, количество связей $l = 3$. Отсюда число степеней свободы составляет $\nu = k - l = 6 - 3 = 3$.

6) С помощью табл. V выясняем, что пороговое значение нашего критерия $\chi^2_{0,05;3} = 7,82$ заметно больше только что полученной величины χ^2. Это значит, что последняя лежит в области приемлемости нулевой гипотезы.

Проведем проверку нулевой гипотезы, исходя из интегральной формы нормального закона. Первые два раздела этой процедуры совпадают с первыми пунктами предыдущей проверки, поэтому переходим сразу к п. 3.

3) С помощью выражения (9.22) вычисляем значения n_i'. Ход расчета показан в столбцах (1) — (11) табл. 9.14; при этом следует помнить, что $\bar{x} = 5,46$, а $\hat{s} = 1,11$.

4) Сравнение эмпирических и теоретических частот по критерию χ^2 показано в столбцах (12) — (14) той же таблицы.

5) Как уже было показано, $\nu = 3$.

6) Пороговое значение χ^2 при $\nu = 3$ и $q = 0,05$ составляет 7,82. Так как полученная нами величина $\chi^2 = 3,66$ меньше этого порога, то она снова находится в области приемлемости гипотезы H_0.

Оценку нашей гипотезы можно было бы провести, как уже говорилось (см. § 4, п. 2), не только через сравнение величин χ^2 и $\chi^2_{q;\nu}$, но также и путем определения вероятности того, что некоторая случайная величина, распределенная по χ^2, примет при $\nu = k - l$ степенях свободы значение, не меньшее, чем вычисленная по формуле (9.17) величина критерия χ^2, т. е. $P(\chi^2_\nu \geqslant \chi^2)$.

Чтобы определить эту вероятность, воспользуемся табл. IX (см. стр. 373, 374), из которой находим, что полученное нами значение χ^2 при трех степенях свободы лежит в интервале между 3 и 4. При этом значение $\chi^2 = 3$ соответствует вероятности 0,3916, а значение $\chi^2 = 4$ — вероятности 0,2615.

Таблица 9.14

	$t_g > x > t_n$	$(t_F)^2 u$	$x - t_g$	$\frac{s}{x - t_g}$	$x - t_n$	$\frac{s}{x - t_n}$	$\Phi\left(\frac{s}{x - t_g}\right)$	$\Phi\left(\frac{s}{x - t_n}\right)$	$\Phi\left(\frac{s}{x - t_n}\right) - \Phi\left(\frac{s}{x - t_g}\right) = {}^t d$	$t d_N = {}^t u$	${}^t d_N - {}^t u$	${}_s(t d_N - t u)$	$\frac{t d_N}{{}_s(t d_N - t u)} = {}_s\chi$	${}^t v$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
1	2,40—3,00	0	—2,46	—2,2153	—3,06	—2,7568	—0,4868	0,4971	0,0103	1,03					1,03
2	3,00—3,60	2	—1,86	—1,6757	—2,46	—2,2153	—0,4535	0,4868	0,0333	3,33	12,63	—2,63	6,9169	0,548	4,36
3	3,60—4,20	8	—1,26	—1,1342	—1,86	—1,6757	—0,3708	0,4535	0,0827	8,27					12,63
4	4,20—4,80	20	—0,66	—0,5946	—1,26	—1,1342	—0,1224	0,3708	0,1484	14,84		5,16	26,6256	1,791	27,47
5	4,80—5,40	21	—0,06	—0,0541	—0,66	—0,5946	—0,0200	0,2224	0,2034	20,24		0,76	0,5776	0,028	47,71
6	5,40—6,00	22	0,54	0,4864	—0,06	—0,0541	0,1879	0,0200	0,2079	20,79		1,21	1,4641	0,070	68,50
7	6,00—6,60	16	1,14	1,0270	0,54	0,4864	0,3461	—0,1879	0,1582	15,82		0,18	0,0324	0,002	84,32
8	6,60—7,20	2	1,74	1,5667	1,14	1,0270	0,4418	—0,3461	0,0957	9,57					93,89
9	7,20—7,80	5	2,34	2,1081	1,74	1,5667	0,4826	—0,4418	0,0408	4,08	15,32	—4,32	18,6624	1,218	97,97
10	7,80—8,40	2	2,94	2,6486	2,34	2,1081	0,4960	—0,4826	0,0134	1,34					99,31
11	8,40—9,00	2	3,54	3,1892	2,94	2,6486	0,4993	—0,4960	0,0033	0,33					99,64
		100							0,9964		99,64			3,657	

Чтобы получить точное значение вероятности, соответствующее вычисленному значению χ^2, произведем интерполяцию:

$$P(\chi^2) = P(3{,}66) = 0{,}3916 - (0{,}3916 - 0{,}2615)\,0{,}66 = 0{,}3916 - 0{,}0859 = 0{,}3057.$$

Эта вероятность не мала (она заметно больше $q = 0{,}05$), следовательно, расхождения между эмпирическими и теоретическими частотами можно считать случайными и несущественными. Иначе говоря, снова оправдывается гипотеза H_0, согласно которой нормальное распределение достаточно хорошо воспроизводит распределение средних длин словоупотреблений в языках мира.

Разумеется, в тех случаях, когда вероятность $P(\chi^2_\nu \geqslant \chi^2)$ явно выше или ниже некоторого заранее заданного уровня значимости $q = 0{,}10$, или $0{,}05$, или $0{,}01$, мы можем принять или отвергнуть гипотезу H_0 без того, чтобы производить утомительные интерполяционные вычисления.

2. Проверка гипотезы о нормальности распределения средних длин словоформ с помощью критерия Колмогорова. Оценка гипотезы о близости теоретического $F(x)$ и выборочного эмпирического $F_N(x)$ распределения может быть осуществлена с помощью **критерия** согласия **Колмогорова**, который использует в качестве меры расхождения между эмпирическим и теоретическим распределениями величину

$$D_N = \max |F_N(x) - F(x)|,$$

с которой мы уже встречались в гл. 8, § 5, п. 2. Нулевая гипотеза H_0 состоит в предположении, что случайная величина X распределена по теоретическому закону $F(x)$, или, иначе говоря, эмпирические значения $F_N(x)$ сходятся по вероятности к теоретическим значениям $F(x)$. Альтернативная же гипотеза H_1 утверждает, что величина X не распределена по $F(x)$.

Если величина X распределена по закону $F(x)$, то можно утверждать (см. гл. 8, § 5, п. 2), что при достаточно большом N с высокой вероятностью $p = K(\lambda)$ выполняется неравенство

$$D_N \sqrt{N} < \lambda, \tag{9.23}$$

которое можно рассматривать как некоторое событие.

Применение критерия согласия Колмогорова представляет нахождение вероятности того, что распределенная по закону Колмогорова случайная величина $D_N \sqrt{N}$ примет некоторое значение, не меньшее, чем λ. В этом случае имеет место неравенство $D_N \sqrt{N} \geqslant \lambda$, являющееся противоположным событием по отношению к неравенству (9.23). По правилу нахождения вероятности противоположного события получаем в случае справедливости гипотезы H_0 следующее равенство:

$$P(D_N \sqrt{N} \geqslant \lambda) = 1 - K(\lambda) = 1 - p = q, \tag{9.24}$$

где $q > 0$ — достаточно малое число.

Итак, вероятность события, состоящего в появлении при единичном испытании неравенства (9.24), очень мала. По существу мы имеем здесь дело с практически неосуществимым событием (ср. гл. 6, § 4, п. 1). Теперь проведем N независимых испытаний, которые в своей сумме можно рассматривать как единичный опыт проверки расхождения распределений $F_N(x)$ и $F(x)$. Если полученная при этом случайная величина $X = D_N \sqrt{N}$ окажется не меньше λ, то это будет означать, что в нашем единичном опыте осуществилось событие, имеющее малую вероятность q. Однако мы предполагали, что такое событие практически неосуществимо в тех случаях, когда справедлива гипотеза H_0, согласно которой эмпирическое распределение $F_N(x)$ сходится к $F(x)$ при $N \to \infty$. Поэтому расхождение

$$D_N = \max |F_N(x) - F(x)|$$

между обоими распределениями следует считать существенным. В итоге мы вынуждены отвергнуть нулевую гипотезу H_0 о том, что случайная величина X имеет распределение $F(x)$.

Напротив, если наш единственный опыт показывает, что $D_N \sqrt{N} < \lambda$, то это означает, что маловероятное событие не осуществилось и у нас нет пока оснований для того чтобы отвергнуть нулевую гипотезу. Такая ситуация свойственна любой научной дисциплине. Чтобы опровергнуть некоторое теоретическое положение, достаточно привести хотя бы один противоречащий пример, а для того чтобы доказать его правильность, примеров уже недостаточно.

Применение критерия Колмогорова осуществляется по следующей схеме.

1. Строится эмпирическая функция $F_N(x)$ и предполагаемая теоретическая функция распределения $F(x)$.

2. Выдвигается гипотеза H_0, согласно которой эмпирическая функция $F_N(x)$ аппроксимирует закон $F(x)$.

3. Ставится опыт, заключающийся в сравнении полученных эмпирических значений $F_N(x_i) = f_i^*$ $(i = 1, 2, \ldots, N)$ с теоретическими значениями $F(x_i) = p_i^*$. Выбирается наибольшая среди них разность D_N^0 и составляется произведение $D_N^0 \sqrt{N} = \lambda_0$.

4. По табл. VII (см. стр. 369, 370) находят вероятность того, что случайная величина $D_N \sqrt{N}$, распределенная по закону Колмогорова, примет значение не меньшее, чем заданное λ_0, т. е.

$$P(D_N \sqrt{N} \geqslant \lambda_0) = 1 - K(\lambda_0) = \gamma.$$

Это и есть вероятность того, что действительное максимальное расхождение $\max |F_N(x) - F(x)|$, объясняемое случайными статистическими флуктуациями, не меньше полученной в нашем опыте максимальной разности D_N^0.

5. Если окажется, что вероятность γ очень мала, т. е. не превышает заранее принятого порога q (обычно берут $q = 0{,}10$; $0{,}05$; $0{,}01$), то исходя из принципа практической невозможности мало-

вероятных событий появление неравенства $D_N\sqrt{N} \geqslant \lambda_0$ считается невозможным событием. Отсюда следует, что расхождение D_N^0 между эмпирическим и теоретическим распределениями нужно считать существенным. Это в свою очередь означает, что гипотеза о близости эмпирического и теоретического распределений неверна.

6. Если же вероятность γ сравнительно велика, т. е. $\gamma > q$, то расхождение D_N^0 следует рассматривать как несущественное, а гипотезу H_0 можно считать совместимой с данными опыта.

Теперь применим критерий Колмогорова к проверке гипотезы о нормальности распределения средних длин словоформ в языках мира.

Таблица 9.15

Интервалы	n_i	$F_N(x)=f_i^*$	n_i'	$F(x)=p_i^*$	$\lvert F_N(x)-F(x)\rvert$
(1)	(2)	(3)	(4)	(5)	(6)
менее 3,00	0	0	1	0,01	0,01
3,00—3,60	2	0,02	3	0,04	0,02
3,60—4,20	8	0,10	8	0,12	0,02
4,20—4,80	20	0,30	15	0,27	0,03
4,80—5,40	21	0,51	20	0,48	0,03
5,40—6,00	22	0,73	21	0,68	0,05
6,00—6,60	16	0,89	16	0,84	0,05
6,60—7,20	2	0,91	10	0,94	0,03
7,20—7,80	5	0,96	4	0,98	0,02
7,80—8,40	2	0,98	1	0,99	0,01
8,40—9,00	2	1,00	1	1,00	0
	100		100		

1) Воспользовавшись данными табл. 9.12, строим эмпирическую функцию $F_N(x)$, представляющую собой накопленную частость $f_i^* = n_i^*/N$, а также теоретическую функцию $F(x)$, являющуюся накопленной вероятностью $p_i^* = n_i^*/N$ [см. столбцы (1) — (5) табл. 9.15].

2) Сформулируем гипотезу H_0, согласно которой функция $F_N(x)$ аппроксимирует закон $F(x)$, представляющий собой интегральную форму нормального распределения.

3) Сравниваем значения $F_N(x) = f_i^*$ и $F(x) = p_i^*$ [столбцы (3) и (5)], в результате чего выбираем наибольшую разность

$$D_N^0 = \max|F_N(x) - F(x)| = 0{,}73 - 0{,}68 = 0{,}05$$

[столбец (6)]; в этом случае имеем $\lambda = 0{,}05\sqrt{100} = 0{,}5$.

4) Из табл. VII находим вероятность того, что случайная величина $D_{100}\sqrt{100}$, распределенная по закону Колмогорова, примет значение, не меньшее, чем 0,5:

$$P(D_{100}\sqrt{100} \geqslant 0{,}5) = 1 - K(0{,}5) = 0{,}9639.$$

5) Полученная вероятность намного больше принимаемого обычно уровня значимости $q = 0{,}05$, поэтому расхождение D_N^0 следует считать несущественным; это дает нам право принять гипотезу о нормальности распределения средних длин словоформ по языкам мира.

При всей своей простоте и удобстве критерий Колмогорова имеет свои недостатки.

Во-первых, он слабо учитывает «хвосты» распределения (малые и большие значения x). Но именно поведение этих «хвостов» оказывается иногда решающим при определении близости эмпирического и теоретического распределений [61, с. 281].

Во-вторых, критерий Колмогорова дает вполне удовлетворительные результаты, когда известен не только вид предполагаемой теоретической функции, но и все ее параметры. Такой случай редко встречается в лингво-статистической практике: обычно параметры теоретического распределения неизвестны, и их приходится оценивать с помощью опытных данных. Применяя критерий χ^2, мы учитывали это обстоятельство путем уменьшения числа степеней свободы. Критерий Колмогорова этой процедуры не предусматривает. Поэтому применение его в случаях, когда параметры теоретического распределения неизвестны, приводит обычно к завышению значения вероятности γ. Это может повлечь за собой принятие гипотезы, плохо согласующейся с опытными данными. Чтобы избежать этой ошибки, некоторые авторы предлагают считать несущественными расхождения между эмпирическим и теоретическим распределением, параметры которого не известны, лишь в том случае, если вероятность $\gamma \geqslant 0{,}6$ [7, с. 271].

3. Проверки гипотезы о нормальности распределения средних длин словоформ с помощью упрощенных критериев. Всякая гипотеза о нормальности вариационного ряда может проверяться не только с помощью таких строгих критериев, как критерий χ^2 Пирсона или критерий Колмогорова, но и оцениваться также с помощью некоторых упрощенных приемов.

Рассмотрение этих приемов начнем с критерия Романовского. Опираясь на правило «трех сигм» (см. гл. 6, § 4, п. 4), В. И. Романовский показал, что

$$P(|\chi^2 - M(\chi^2)| < 3\sigma_{\chi^2}) \geqslant 0{,}98, \quad \text{когда} \quad \nu \leqslant 7,$$

$$P(|\chi^2 - M(\chi^2)| < 3\sigma_{\chi^2}) \geqslant 0{,}99, \quad \text{когда} \quad \nu > 7,$$

где ν — число степеней свободы. Учитывая, что при больших ν распределение χ^2 асимптотически приближается к нормальному распределению с параметрами $M(\chi^2) = \nu$, $D(\chi^2) = \sigma_{\chi^2}^2 = 2\nu$ ($\sigma_{\chi^2} = \sqrt{2\nu}$) (см. гл. 8, § 3, п. 1), и опираясь на принцип практической уверенности (см. гл. 6, § 4, п. 1), нетрудно прийти к простому правилу, согласно которому, если

$$R_m = |\chi^2 - \nu| / \sqrt{2\nu} < 3,$$

то расхождение между эмпирическим и нормальным распределениями можно считать с вероятностью 0,98 и 0,99 случайным; если же

$$R_m = |\chi^2 - \nu| / \sqrt{2\nu} \geqslant 3,$$

то расхождение следует считать существенным.

Пользуясь этими правилами, проверим гипотезу о нормальности распределения средних длин словоупотреблений в языках мира.

Так как $\chi^2 = 3{,}66$ (см. табл. 9.14), а $\nu = 3$, то

$$R_m = \frac{|3{,}66-3|}{\sqrt{2\cdot 3}} = \frac{0{,}66}{\sqrt{6}} = \frac{0{,}66}{2{,}45} \approx 0{,}27.$$

Здесь $R_m < 3$, поэтому можно считать, что расхождение между эмпирическим и теоретическим распределением случайно и нормальное распределение достаточно хорошо воспроизводит интересующее нас эмпирическое распределение.

Нормальность вариационного ряда может быть проверена с помощью *чисел Вестергарда*. Эта проверка, опирающаяся на правила «двух» и «трех сигм», строится по следующей схеме.

Сначала задаются четыре множителя 0,3; 0,7; 1,1; 3, затем вычисляется средняя арифметическая $\bar{x} = \bar{F}$ и среднее квадратическое отклонение σ. Гипотеза о нормальности рассматриваемого эмпирического распределения не отвергается, если удовлетворяются следующие условия:

а) в промежутке от $\bar{x} - 0{,}3\sigma$ до $x + 0{,}3\sigma$ находится не менее 0,25 всей совокупности;

б) в промежутке от $\bar{x} - 0{,}7\sigma$ до $\bar{x} + 0{,}7\sigma$ расположено не менее 0,50 всей совокупности;

в) в промежутке от $\bar{x} - 1{,}1\ \sigma$ до $\bar{x} + 1{,}1\sigma$ умещается не менее 0,75 совокупности;

г) в промежутке от $\bar{x} - 3\sigma$ до $\bar{x} + 3\sigma$ находится не менее 0,998 всей совокупности.

Теперь, используя числа Вестергарда, еще раз проверим гипотезу о нормальности распределения средних длин словоупотреблений в языках мира.

Учитывая, что $\bar{x} = 5{,}46$, а $\sigma = \hat{s} = 1{,}11$ (эти значения были найдены в п. 1 при проверке рассматриваемой гипотезы с помощью критерия χ^2 Пирсона), сгруппируем данные табл. 7.28 (см. стр. 260) так, как этого требуют условия Вестергарда (табл. 9.16).

Эта группировка показывает [см. столбец (3)], что распределение средних длин словоформ подчиняется нормальному закону.

Проверку нормальности эмпирического распределения можно осуществить графическим путем с помощью *вариационной сетки Турбина* [7, с. 249], представляющей прямоугольную систему ко-

Таблица 9.16

Интервалы	Количество языков, попадающих в интервал	Доля совокупности, попадающая в интервал
(1)	(2)	(3)
5,13—5,80	27	0,27
4,68—6,24	58	0,58
4,24—6,68	77	0,77
2,13—8,79	100	1,00

ординат, в которой по оси абсцисс откладывается линейный масштаб*, а по оси ординат наносится логарифмическая шкала, соответствующая интегральной функции

$$F(x) = \frac{1}{2} + \Phi(x) = \frac{1}{2\pi} \int_{-\infty}^{x} e^{-z^2/2}\, dz.$$

Шкала на оси ординат состоит из разных отрезков—циклов, каждый из которых соответствует изменению характеристики логарифма на единицу. Циклы разбиваются на части соответственно изменению мантисс десятичных логарифмов. На логарифмической шкале не может быть нулевой отметки, поскольку $\lg 0 = -\infty$.

Идея использования сетки Турбина заключается в следующем.

Пусть имеется система координат x, y ($0 \leqslant y \leqslant 1$) с линейной шкалой по оси x и нелинейной — по оси y. Путем некоторого преобразования $y \to z = \varphi(x)$ получается система координат x, z с линейной шкалой по оси ординат. Для этого функцию $F(x)$ нормального распределения с параметрами μ и σ^2 можно нормировать с помощью линейного преобразования $z = (x - \mu)/\sigma$ (ср. гл. 8). Интегральная кривая нормального распределения $y = F(x)$ асимптотически приближается к нулю и к единице (т. е. к 100%; рис. 65). Поэтому наша шкала по оси ординат не может достичь ни нуля, ни единицы. При переходе к функции $z = \varphi(x)$ интегральная кривая превращается в бесконечную прямую, имеющую наклон $1/\sigma$ и проходящую через точку $\bar{x} = \mu$. При построении графика удобно пользоваться характеристическими значениями, приведенными в табл. 9.17.

Таблица 9.17

x	$\mu-3\sigma$	$\mu-2\sigma$	$\mu-\sigma$	μ	$\mu+\sigma$	$\mu+2\sigma$	$\mu+3\sigma$
z	—3	—2	—1	0	1	2	3
y (в %)	0,14	2,28	15,87	50,00	84,13	97,72	99,86

* Иногда на ось абсцисс наносится логарифмическая шкала.

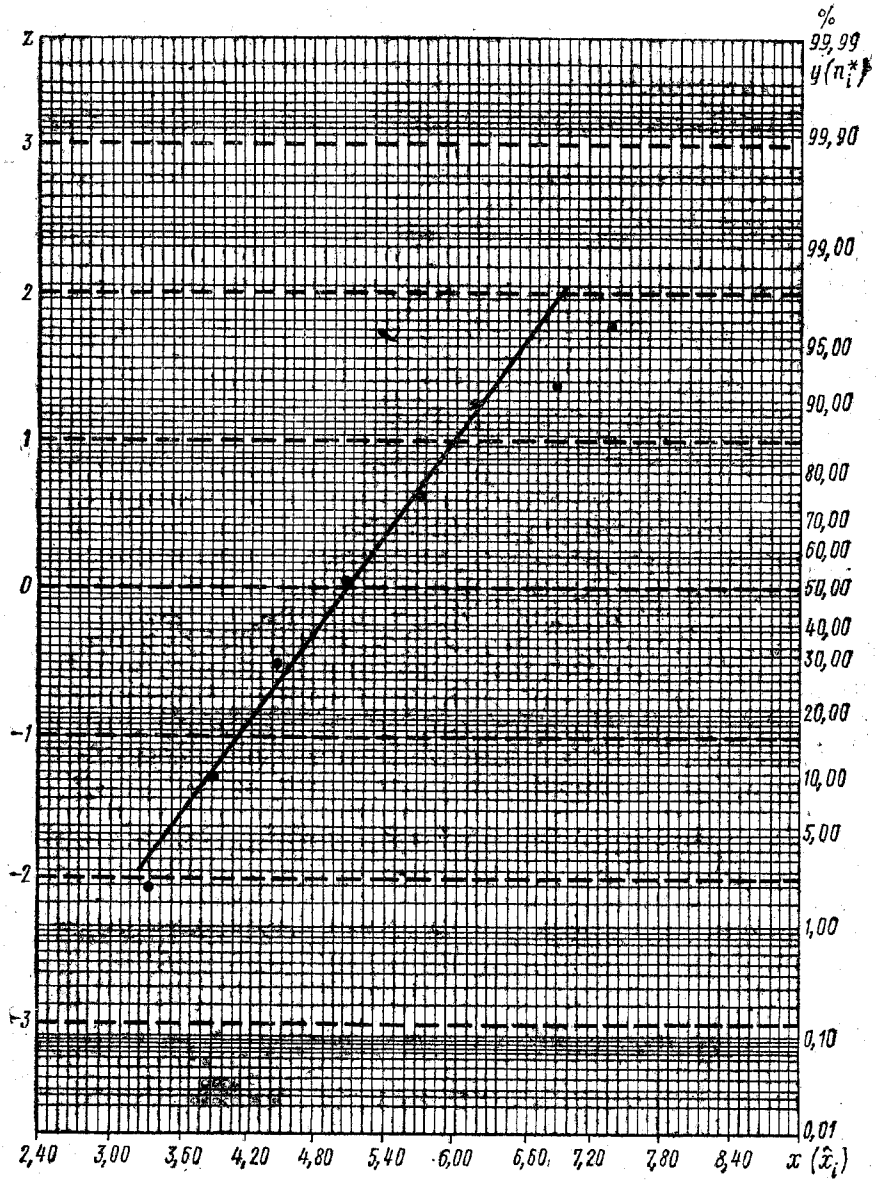


Рис. 65

Схема использования сетки Турбина для проверки нормальности лингвистического распределения предусматривает следующие операции.

1. По оси абсцисс откладываются значения признака, взятые из эмпирического лингвистического распределения (вариационного ряда), а по оси ординат—соответствующие им накопленные частости (f^*).

2. Если эмпирическое распределение хорошо аппроксимируется нормальным распределением, то его интегральная кривая образует на «вероятностной бумаге» в интервале ординат 10 — 90% прямую линию. Если же эмпирическое распределение не укладывается в нормальное распределение, то точки, соответствующие накопленным частостям f^*, заметно отходят от прямой линии.

Для оценки гипотезы о нормальности распределения средних длин словоупотреблений в языках мира нанесем на сетку Турбина (рис. 65) накопленные частоты n_i^* [см. табл. 9.14, столбец (15)], соответствующие серединам x_i интервалов. Полученные точки в интервале 2 — 90% располагаются близко к прямой. Это снова говорит о том, что распределение средних длин словоупотребления в языках мира близко к нормальному*.

Итак, мы проверили гипотезу о нормальности распределения средних длин словоформ в различных языках мира с помощью шести статистических критериев. Все эти критерии подтвердили справедливость указанной гипотезы.

4. Проверка с помощью критерия Колмогорова — Смирнова гипотезы о согласии распределений средних длин словоформ в двух языковых семьях. Используя различные статистические критерии, мы выяснили, что средние длины словоформ в языках мира распределены нормально. Отсюда можно предположить, что распределения средних длин словоформ в разных семьях языков идентичны. Проверим это предположение применительно к финноугорским и тюркским языкам с помощью **критерия Колмогорова — Смирнова,** основные идеи которого сводятся к следующему. Пусть эмпирическая интегральная функция распределения $F_{N_1}(x)$ построена по выборке $x_1, x_2, \ldots, x_{N_1}$. Одновременно имеется другая интегральная функция $F_{N_2}(x)$, построенная по выборке $x_1, x_2, \ldots, x_{N_2}$. Если объемы выборок N_1 и N_2 неограниченно возрастают так, что отношение N_2/N_1 остается постоянным, то при условии, что $N_0 = N_1N_2/(N_1 + N_2) \to \infty$ вероятность неравенства

$$\{D_{N_1 N_2} = \max |F_{N_1}(x) - F_{N_2}(x)|\} < \lambda/\sqrt{N_0}$$

стремится к функции Колмогорова (см. гл. 8, § 5, п. 2)

$$K(\lambda) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2\lambda^2},$$

* Ср. использование сетки Турбина в работе [33, с. 206—243] при определении нормальности вариационных рядов стилистических признаков, использующихся при установлении авторства анонимного текста.

где $\lambda > 0$; иными словами,

$$P(D_{N_1N_2}\sqrt{N_0} < \lambda) \approx K(\lambda) = \mathfrak{p}. \qquad (9.25)$$

Применение критерия Колмогорова — Смирнова, так же как и использование критерия Колмогорова, построено на нахождении вероятности того, что величина $D_{N_1N_2}\sqrt{N_0}$ примет некоторое значение, не меньшее, чем λ. В этом случае имеет место неравенство

$$D_{N_1N_2}\sqrt{N_0} \geqslant \lambda, \qquad (9.26)$$

которое является противоположным событием по отношению к неравенству, входящему в выражение (9.25). По правилу нахождения вероятности противоположного события получаем

$$P(D_{N_1N_2}\sqrt{N_0} \geqslant \lambda) = 1 - K(\lambda) = 1 - \mathfrak{p} = \mathfrak{q}, \qquad (9.27)$$

где $\mathfrak{q} > 0$ — достаточно малое число.

Выражения (9.26) и (9.27) используются для проверки нулевой гипотезы H_0, состоящей в предположении, что оба эмпирических распределения, находясь в полном согласии [т. е. $F_{N_1}(x) = F_{N_2}(x)$], представляют выборки N_1 и N_2, принадлежащие одной и той же генеральной совокупности. Весь ход рассуждений при проверке H_0 аналогичен рассуждениям, на которых строилось использование критерия Колмогорова (п. 2), поэтому повторять их здесь мы не будем, а ограничимся описанием общей схемы применения критерия Колмогорова—Смирнова, которая включает следующие операции:

1. По наблюдаемым значениям выборок N_1 и N_2 составляются эмпирические интегральные функции распределения $F_{N_1}(x)$ и $F_{N_2}(x)$.

2. Выдвигается гипотеза H_0 о полном согласии функций $F_{N_1}(x)$ и $F_{N_2}(x)$ и принадлежности выборок N_1 и N_2 к одной генеральной совокупности.

3. Ставится опыт по сравнению величин $F_{N_1}(x_l)$ и $F_{N_2}(x_l)$. Выбирается наибольшая разность $D^0_{N_1N_2}$ и составляется произведение

$$D^0_{N_1N_2}\sqrt{N_0} = D^0_{N_1N_2}\sqrt{\frac{N_1N_2}{N_1+N_2}} = \lambda_0.$$

4. Из табл. VII на стр. 369, 370 определяется вероятность γ того, что случайная величина $D_{N_1N_2}\sqrt{N_0} = \lambda$, имеющая распределение Колмогорова, примет значение, не меньшее, чем λ_0. Величина γ есть одновременно вероятность того, что действительное максимальное расхождение

$$D_{N_1N_2} = \max |F_{N_1}(x) - F_{N_2}(x)|,$$

которое можно отнести за счет случайных статистических флуктуаций, не меньше полученной в нашем опыте максимальной разности $D^0_{N_1N_2}$.

5. Если вероятность γ очень мала, т. е. не выше заданного порога $\mathfrak{q}$ (обычно $\mathfrak{q} = 0{,}10; 0{,}05; 0{,}01$), то, опираясь на принцип практической невозможности маловероятных событий, мы должны считать появление неравенства

$$\{D_{N_1N_2}\sqrt{N_0} = \lambda\} \geqslant \lambda_0$$

невозможным событием. При этом расхождение $D^0_{N_1N_2}$ между нашими эмпирическими расхождениями следует рассматривать как существенное. Отсюда следует, что гипотеза H_0 о согласии обоих распределений и принадлежности выборок N_1 и N_2 к одной генеральной совокупности должна быть отвергнута.

6. Если же вероятность γ сравнительно велика, т. е. больше заданной вероятности $\mathfrak{q}$, то разность $D^0_{N_1N_2}$ можно считать несущественной. Поэтому следует принять гипотезу H_0 о согласии распределений $F_{N_1}(x)$ и $F_{N_2}(x)$ и принадлежности выборок N_1 и N_2 к одной генеральной совокупности.

Особенность критерия Колмогорова—Смирнова состоит в том, что он со сколь угодно большой вероятностью позволяет обнаруживать любое расхождение между двумя эмпирическими функциями $F_{N_1}(x)$ и $F_{N_2}(x)$ при условии, что N_1 и N_2 достаточно велики. Этот критерий применяется тогда, когда нужно проверить полное согласие обоих распределений на всем интервале изменения случайной величины X и когда для этой проверки имеется очень обширный материал наблюдений.

Применяя приведенную выше схему, проверим гипотезу о согласии распределений средних длин словоформ в тюркских и финно-угорских языках.

1) Воспользовавшись данными табл. 7.27 (см. стр. 259), составим вспомогательную таблицу (табл. 9.18), в столбцах (6) и (7) которой приведены эмпирические интегральные функции распределения длин словоформ: $F_{N_1}(x)$ для тюркских и $F_{N_2}(x)$ для финноугорских языков.

2) Выдвигаем гипотезу H_0 о согласии этих функций и принадлежности выборок N_1 (тюркские тексты) и N_2 (финноугорские тексты) к одной генеральной совокупности с точки зрения распределения длин словоформ.

3) Выбираем наибольшую разность

$$D^0_{N_1N_2} = \max |F_{N_1}(x) - F_{N_2}(x)| = |0{,}278 - 0{,}714| = 0{,}436$$

и определяем пороговое значение

$$\lambda_0 = 0{,}436\sqrt{18 \cdot 14/(18+14)} = 0{,}436\sqrt{7{,}875} = 1{,}223.$$

4) Пользуясь табл. VII, выясняем, что случайная величина λ, имеющая распределение Колмогорова, примет с вероятностью 0,10 значение, не меньшее, чем $\lambda_0 = 1{,}223$.

5) Полученная вероятность не мала, поэтому можно принять нулевую гипотезу, которая утверждает, что с точки зрения распределения средних длин словоформ тюркские и финноугорские тексты принадлежат к одной генеральной совокупности, в качестве которой, как это было показано в п. 1 — 3 § 5, выступают тексты языков мира.

Таблица 9.18

Длина слов в буквах	Частота слов с данной длиной		Накопленная частота слов с данной длиной		Относительно накопленная частота слов с данной длиной		$F_{N_1}(x)-F_{N_2}(x)$
	тюркские языки F_1	финно-угорские языки F_2	тюркские языки F_2^*	финно-угорские языки F_2^*	тюркские языки $F_{N_1}(x)=f_1^*$	финно-угорские языки $F_{N_2}(x)=f_2^*$	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
4,1	0	1	0	1	0	0,071	0,071
4,7	0	1	0	2	0	0,143	0,143
4,8	0	2	0	4	0	0,286	0,286
5,1	1	1	1	5	0,056	0,357	0,301
5,2	1	0	2	5	0,111	0,357	0,246
5,3	0	2	2	7	0,111	0,500	0,389
5,4	1	1	3	8	0,167	0,571	0,404
5,5	2	2	5	10	0,278	0,714	0,436
5,6	3	0	8	10	0,445	0,714	0,269
5,8	4	1	12	11	0,667	0,785	0,118
5,9	1	0	13	11	0,723	0,785	0,062
6,0	1	0	14	11	0,778	0,785	0,007
6,2	0	1	14	12	0,778	0,857	0,079
6,3	2	0	16	12	0,889	0,857	0,032
7,4	0	1	16	13	0,889	0,928	0,039
7,5	1	0	17	13	0,944	0,928	0,061
7,8	0	1	17	14	0,944	1,000	0,056
8,8	1	0	18	14	1,000	1,000	0,000
	18	14					

Этот, казалось бы, скромный статистический результат может иметь важные теоретические последствия не только в области теории языка или психолингвистики, но и в плане кибернетической физиологии высшей нервной деятельности, а также инженерной лингвистики. Действительно, нормальность распределения длин словоформ может рассматриваться как указание на то, что существует некоторый общечеловеческий эталон, равный центру этого распределения. Величину этого эталона, определяющуюся, вероятно, особенностями строения быстродействующей памяти человека [21, с. 97], следует учитывать при расчете памяти слушающих, переводящих и обучающих автоматов.

Существование некоторого среднего общечеловеческого эталона длины словоформы и нормального распределения средних длин вокруг этого эталона в языках мира нисколько не противоречит тому факту, что длина словоформы может неограниченно возрастать и в отдельных конкретных случаях достигает нескольких десятков букв (см. гл. 4, § 2, п. 1). Легко заметить, что искусственные слова типа *двустороннесимметричнообразный* «не выговариваются» сразу, «одним дыханием», а делятся при воспроизведении на сегменты, включающие одну или более морфем, причем длина этих сегментов

приближается к средней длине словоформы. Аналогичным образом осуществляется, очевидно, и восприятие сверхдлинных слов.

Нет ничего необычного и в том, что в некоторых стилях средняя длина словоформы может заметно возрастать по сравнению с общечеловеческим эталоном или средней длиной, характерной для данного языка (ср. § 3, п. 1). Эти отклонения свидетельствуют лишь о существовании у нашей быстродействующей памяти некоторых резервов, позволяющих человеку пользоваться различными функциональными разновидностями языка.

§ 6. Доминантные смысловые единицы и элементы заполнения текста

1. Исследование вариационных рядов словосочетаний в немецких публицистических текстах. Статистическая проверка лингвистических гипотез используется для формально-семантического исследования текста и составляющих его единиц.

Рассмотрим в этой связи распределение трехкомпонентных немецких словосочетаний, образованных от наиболее частых словоформ. Поскольку трехсловные сочетания являются редкими лингвистическими событиями, выдвигается нулевая гипотеза H_0, согласно которой их распределения подчиняются закону Пуассона.

Проверка этой гипотезы осуществляется с помощью критерия χ^2 и проходит по схеме, близкой к алгоритму проверки нормальности вариационного ряда (ср. § 5).

1. Сначала находят вероятность появления частот $F = 0, 1, 2$ и т. д. по известной формуле распределения Пуассона

$$p_t = P_N(F) \approx \frac{\lambda^F}{F!} e^{-\lambda}, \qquad (9.28)$$

где λ — параметр распределения. Поскольку этот параметр обычно неизвестен, его оценивают с помощью средней $\bar{x} = \bar{F}$ эмпирического распределения.

2. Определяют теоретические частоты $Np_t = S_F^T$, которые по формуле (9.17) сравнивают с опытными частотами $n_t = S_F$. В тех случаях, когда величины S_F^T и S_F меньше пяти, они соответственно объединяются с соседними значениями (ср. табл. 9.13).

3. Полученную величину χ^2 сопоставляют с пороговым значением $\chi^2_{q;\nu}$, выбранным из табл. V (см. стр. 366), исходя из заданного уровня значимости q и числа степеней свободы, равного $\nu = N^* - 2$. Принятие или непринятие гипотезы H_0 о пуассоновском характере рассматриваемого распределения зависит от того, выполняется ли неравенство $\chi^2 < \chi^2_{q;\nu}$ или $\chi^2_{q;\nu} \leqslant \chi^2_{q;\nu}$.

4. Оценку гипотезы H_0 можно осуществить также путем определения вероятности того, что некоторая случайная величина, распределенная по χ^2, примет при $\nu = N^* - 2$ степенях свободы значение, не меньшее, чем вычисленная по формуле (9.17) величина критерия χ^2 (см. § 4, п. 2).

По указанной схеме исследовались распределения у большой группы различных по грамматической природе и семантике трехсловных сегментов, взятых из немецких газетных текстов [33, с. 131 — 162]. Ход исследования проследим на примере сегмента Δ daβ der (начало дополнительного придаточного предложения с существительным в начальной позиции, символ Δ указывает здесь на знак препинания), распределение которого уже рассматривалось нами в гл. 6 (см. § 3, п. 2).

1) Имеется 100 серий (S), каждая из которых содержит 1000 трехсловных сочетаний. По формуле (9.28) определяем вероятность $P_N(F)$ ($F = 0, 1, 2$ и т. д.) появления сегмента Δ daβ der в каждой серии [столбец (3) табл. 9.19]. Вместо параметра λ используем его оценку

$$\overline{F} = \frac{1}{100}(0\cdot 32 + 1\cdot 38 + 2\cdot 19 + 3\cdot 5 + 4\cdot 6) = \frac{115}{100} = 1{,}15$$

[в связи с укрупнением последних частот — ср. столбцы (2) и (4), полученная оценка несколько отличается от $\overline{F}$, вычисленного на основании данных табл. 6.7].

2) Затем вычисляем теоретические частоты S_F^T [столбец (4)], которые сравниваем с эмпирическими частотами S_F [столбцы (5) и (7)].

3) Полученная в столбце (7) величина $\chi^2 = 4{,}67$ меньше порогового значения $\chi^2_{0,05;\,3} = 7{,}82$ (с учетом укрупнения малых частот имеем $\nu = 5 - 2 = 3$) и лежит таким образом в области принятия нулевой гипотезы о пуассоновском характере интересующего нас лингвистического распределения.

4) Оценивая гипотезу H_0 с помощью вероятности $P(\chi^2)$, мы убеждаемся, что вероятность некоторой распределенной по χ^2 случайной величины принять значение, большее, чем 4,67 при $\nu = 3$, лежит в пределах между 0,26 и 0,17 (см. табл. IX). Иными словами, эта вероятность достаточно велика. Поэтому гипотеза о пуассоновском характере распределения сегмента Δ daβ der вполне правдоподобна.

Таблица 9.19

Номер интервала	F	S_F	$P_N(F)$	S_F^T	$S_F - S_F^T$	$(S_F - S_F^T)^2$	$\frac{(S_F - S_F^T)^2}{S_F^T}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	0	32	0,3103	31,03	0,97	0,9216	0,03
2	1	38	0,3628	36,28	1,69	2,8561	0,08
3	2	19	0,2123	21,23	−2,24	5,0176	0,23
4	3	5	0,0828	8,28	−3,28	10,7584	1,30
5	4	4 } 6	0,0242	2,99	3,01	9,0601	3,03
	5	2 }	0,0057				
		100	0,9981				$\chi^2 = 4{,}67$

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

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

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

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

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

2. Статистическое опознание термина и выделение доминантных единиц в тексте. Гипотеза о возможности формально-статистического опознания термина была рассмотрена относительно пуассоновского, нормального и логнормального распределений [32а, с. 47—112]. Исследованию подвергались вариационные ряды двух единиц искусственных языков (буквенные символы, цифры), 298 словоформ и 300 трехсловных сегментов, взятых из английских научно-технических текстов подъязыка судовых механизмов. Общий объем

исследованного текста составил 400 тыс. словоупотреблений (ок. 1200 стр.).

Применяя электронно-вычислительную технику, авторы рассмотрели разные варианты разбивки и нормировки материала, в результате чего для каждой словоформы и словосочетания были построены до 27 вариационных рядов. Затем с помощью ЭВМ по критерию χ^2 при уровне значимости 0,01 проверялось согласие этих рядов с указанными выше распределениями.

Поскольку учесть все совпадения и несовпадения эмпирических и теоретических данных в 27 вариационных рядах каждой лексической единицы невозможно, было введено следующее обобщающее условие. Эмпирическое распределение считалось совпадающим с теоретическим в том случае, если такое согласие показывал вариационный ряд, охватывающий целиком всю выборку, либо несколько вариационных рядов, построенных на частных выборках, составляющих в сумме весь выборочный текст длиной в 400 тыс. словоупотреблений. В других случаях отмечалось несогласие эмпирического и теоретического распределений.

Исследование показало, что согласие или несогласие эмпирического распределения с тем или иным теоретическим законом зависит, во-первых, от статистических условий эксперимента — частоты лингвистической единицы и способа построения вариационного ряда, а во-вторых, от семантики исследуемой словоформы или словосочетания.

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

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

2. В зоне средних частот (номера 60 — 1500 списка) отмечается постепенное отступление нормального закона перед законом Пуассона. При этом пуассоновское распределение появляется в первую очередь в тех вариационных рядах словоформ, которые обобщают малые микровыборки (1 — 2 тыс. словоупотреблений), наоборот, нормальность удерживается в рядах, построенных из крупных серий по 4,8 или 16 тыс. словоупотреблений. Расширяя сферу своего распространения, пуассоновские распределения накладываются на нормальные: начиная с 60-70-х номеров списка все чаще попадаются словоформы, вариационные ряды которых при микровыборках объемом в 2 — 16 тыс. словоупотреблений подчиняются и нормальному, и пуассоновскому законам. Сходная картина наблюдается и в частотном списке словосочетаний с той лишь разницей, что наложение пуассоновского распределения на нормальное обнаруживается в самом начале списка.

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

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

Обращаясь к семантическому анализу полученных результатов, разобьем все лексические единицы на две группы. В первую группу войдут словоформы и словосочетания, эмпирические распределения которых не показывают согласия ни с нормальным законом, ни с законом Пуассона. Вторая группа будет включать лексические единицы, вариационные ряды которых подчиняются либо нормальному, либо пуассоновскому распределению, либо одновременно обоим законам. Такая разбивка экспериментального материала оправдывается тем, что при больших значениях параметра λ распределение Пуассона приобретает форму нормального распределения (см. гл. 6, § 3, п. 2). Это аппроксимирующее свойство нормального закона обнаруживается, в частности, в наложении этого распределения на закон Пуассона в зоне среднечастотных словоформ и словосочетаний, о чем говорилось выше.

Обратимся к анализу первой группы. Просмотр приведенных в указанной выше работе лексических списков [32а, с. 82 — 88, 92 — 97] показывает, что из 298 словоформ эмпирические распределения 66 единиц не подчиняются ни нормальному, ни пуассоновскому законам, а среди 300 словосочетаний 22 не дают согласия своих вариационных рядов с указанными теоретическими законами.

Какова же лексико-грамматическая природа этих слов и словосочетаний?

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

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

the gas turbine 'газовая турбина', marine disel engine 'судовой дизельный двигатель'; the water level 'уровень воды'; и др.

current 'текущий'; moving 'приводя(щий) в движение'; marine 'морской'; work 'работать' и т. п.

Во-вторых, не подчиняется нормальному и пуассоновскому закону эмпирическое распределение вспомогательных и модельных

глагольных форм was, would, should, must, а также распределение глагольного оборота shown in fig ... 'показано на рис. ...'.

В-третьих, в рассматриваемую группу попадают сегменты текста, представляющие собой комбинацию словоформ и символов искусственных языков — цифр и формул (x), а также буквенных обозначений и сокращений (z). Ср.: x and x 'x и x'; is xz 'равняется xz'; figure xz 'рисунок xz'; see zx' см. zx'; value of x 'значение x', и т. п.

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

Метод проверок гипотезы о согласии эмпирического и теоретического распределения с помощью критерия χ^2 применялся не только к научно-техническим текстам, но и к художественной прозе, публицистике, а также к речи душевнобольных. При этом выяснилось, что несогласие эмпирического распределения с нормальным и пуассоновским законами обнаруживают не только терминологические слова и словосочетания, а также близкие к ним по семантической природе комбинации словоупотреблений и элементов искусственных языков, но и другие лексические единицы, обладающие важным, с точки зрения сюжета текста, значением. В художественной речи этим свойством обладают слова и словосочетания, выражающие основные образы произведения, в публицистике — имена собственные и названия особо важных политических, экономико-социальных и географических реалий (см. § 3, п. 3), в речи душевнобольных—слова и словосочетания, обозначающие объекты навязчивого состояния. Короче говоря, несогласие эмпирических и теоретических распределений обнаруживают семантически доминантные (ключевые) единицы текста, напротив, недоминантные единицы текста, обладающие «стертой» семантикой, показывают хорошее согласие своих распределений с теоретическими законами Пуассона и Гаусса.

Природа этого явления состоит, очевидно, в следующем.

Как уже говорилось (см. гл. 5), порождение текста определяется, с одной стороны, системой языка и его нормой, а с другой — независимой от языка ситуацией, которую призван описывать этот текст. Появление семантически нагруженных ключевых элементов текста диктуется ситуацией, употребление же семантически «стертых» единиц, так называемых *единиц заполнения*, служащих средством организации текста и связи ключевых единиц, подчиняется требованиям системы и нормы языка. Отсюда следует, что статистика единиц заполнения подчиняется вероятностным законам нормы языка, в то время как статистика ключевых элементов текстов управляется вероятностью ситуаций.

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

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

Нормальный закон и закон Пуассона хорошо описывают распределения служебных слов и общеупотребительных слов и словосочетаний, в частности, потому, что этим последним присуще сравнительно независимое употребление в тексте. Напротив, лингвистические единицы, характеризующиеся сильными валентностями, статистико-вероятностным выражением которых являются марковские связи текста, дают, как правило, скошенные вправо эмпирические распределения, имеющие иногда несколько вершин. Эти вариационные ряды не подчиняются законам Пуассона и Гаусса, но зато могут описываться логнормальным распределением (см. гл. 6, § 3, п. 5) или кривыми Пирсона с правой асимметрией [32 в, с. 335—361]. Вариационные ряды, описываемые логнормальным распределением, дают уже упоминавшиеся английские вспомогательные и модальные глаголы was, would, should, must. Хотя эти глагольные формы не являются ключевыми единицами текста, они, очевидно, тесно связаны с такими знаменательными словоформами, которые несут в тексте смысловую или эмоционально-выделительную нагрузку.

Описанные особенности распределений служебных и общеупотребительных лексических единиц, а также терминологических и вообще доминантных элементов текста могут быть использованы в качестве диагносцирующего аппарата, автоматически распознающего в тексте ключевые слова и выражения текста. Извлечение с помощью ЭВМ из больших массивов текста семантически нагруженных ключевых единиц позволяет, с одной стороны, автоматизировать трудоемкие процессы составления алгоритмов машинного реферирования и семантического перевода текста. С другой стороны, статистико-автоматическое исследование семантики текста на машине дает богатый материал для решения таких теоретических и прикладных задач, как исследование соотношений «смысл — текст», «норма — речь», построение обучающих алгоритмов, лингвистическая диагностика душевных заболеваний и других вопросов.

ЗАКЛЮЧЕНИЕ

1. Краткий обзор содержания. Конфронтация языка и математики осуществлялась нами, с одной стороны, в области количественной экспликации диахронических процессов и процессов развертывания текста, а с другой — в плане вероятностно-статистического моделирования построения текста, являющегося результатом взаимодействия системы, нормы и ситуации.

Одна из задач этой конфронтации состояла в выявлении таких математических моделей, с помощью которых можно было бы не только описать структуру и функционирование интересующих нас лингвистических объектов, но и получить новую информацию о природе языка и речи.

Экспликация синтагматического развертывания речи и диахронии языка с помощью аппарата математического анализа (см. гл. 1 — 4) обнаружила, что лингвистические процессы имеют ступенчатый (скачкообразный) и циклический характер. Цикличное распределение информации в тексте диктуется периодическим ритмом, в котором мозг перерабатывает поступающую в него информацию.

Ступенчатость и цикличность развития языка определяется, очевидно, дискретно-порционным характером перестройки языковой системы.

Общность математических моделей, аппроксимирующих развитие языка и информационное развертывание текста, заставляет лингвиста задуматься над проблемой речевого онтогенеза и диахронического филогенеза.

Экспликация построения текста с помощью аппарата теорий вероятностей и информации, а также математической статистики (см. гл. 5 — 9) показала, что порождение текста определяется не только системой языка, но диктуется также, с одной стороны, вероятностными моделями, а с другой — вероятностью. Поэтому информационно-статистическая структура текста является результатом взаимодействия двух статистик. Одной из них является статистика нормативных единиц заполнения (буквы, слоги, грамматические морфемы, служебные слова и словосочетания, а также некоторые синтаксические схемы). Другой статистикой является статистика доминантных, или ключевых, единиц текста (слов, словосочетаний и синтаксических построений), передающих основное содержание текста. Несогласованность этих статистик может быть использована для формального автоматического распознавания терминологических единиц текста.

Одновременно возникает вопрос о том, как взаимодействуют между собой элементы заполнения и ключевые единицы в рамках циклической информационной схемы текста — схемы, навязанной ритмом работы нашего мозга. Исследование этого вопроса является одной из важных задач современной кибернетической лингвистики.

2. Перспективы развития математической лингвистики. Осуществленная в книге конфронтация языка и математики проходила под знаком преобладания математического подхода к отбору материала и его компоновке. Такой подход оказался целесообразным ввиду того, что по сравнению с языкознанием математика имеет более строгую и последовательную организацию. Однако при этом подходе в сферу математической лингвистики включаются лишь те явления языка и речи, которые могут быть подвергнуты экспликации и моделированию с помощью жесткого аппарата современной «количественной» и «качественной» математики, охватывающей не только функции одной переменной, рассмотренные в настоящей книге, но и функции многих переменных (случайных величин). Вместе с тем процесс математизации лингвистики вовсе не направлен на обеднение языкознания и его подчинение математике. Напротив, этот процесс должен привести к усилению исследовательского аппарата языкознания и к обогащению наших лингвистических знаний.

Поэтому вслед за этапом конфронтации математики и языкознания — этапом, на котором математика исполняла роль «королевы наук», должен следовать этап сближения, на котором математика выступает на службе остальных наук, создавая для языкознания и для других гуманитарных наук особый логический аппарат.

Необходимость в создании такого аппарата объясняется тем, что используемый в настоящее время для моделирования и экспликации лингвистических явлений традиционный математический аппарат был первоначально предназначен для описания «жестких» и сравнительно простых систем неживой природы. Поэтому он оказывается недостаточно адекватным при моделировании сложных гуманистических, в том числе языковых систем, имеющих, как принято сейчас говорить, «мягкую», полиморфную структуру [26, с. 208]. Эта неадекватность перерастает иногда в несовместимость традиционного математического аппарата и сложных гуманистических систем. Суть этой несовместимости кратко определяется так: чем сложнее система, тем менее способна традиционная математика дать точные и одновременно имеющие практическое значение суждения о поведении этой системы [65, с. 7].

Преодолеть этот парадокс можно с помощью такого математического аппарата, который использовал бы эвристические подходы и приемы, при помощи которых человек решает различные жизненные и в том числе лингвистические задачи.

Создавая такой аппарат, следует иметь в виду, что оперативными единицами человеческого мышления являются не дискретные математические объекты, а элементы некоторых нечетких множеств (гл. 1, § 1, п. 1). При этом переход от принадлежности элементов $x_1, x_2, \ldots, x_i$ нечеткому множеству A к «непринадлежности» этих элементов тому же множеству осуществляется не скачкообразно, а непрерывно, характеризуясь убыванием степени принадлежности μ элемента x_i множеству A в интервале от единицы до нуля. Сим-

волически это записывается так:

$$1 \geqslant \mu\,(x_i \in A) \geqslant 0.$$

Использование нечетких множеств для описания систем лингвистических объектов имеет принципиальное методологическое значение для языкознания. Задавая каждому элементу лингвистического множества коэффициент принадлежности μ, мы получаем возможность математически описать размытость границ семантических полей, слов, словосочетаний и предложений, акустических признаков фонемы, границ диалектов и говоров, а также эксплицировать парадокс между коллективным и индивидуальным владением языка (парадокс языка и идиолекта). С этим парадоксом мы неоднократно встречались в книге, рассматривая субъективные вероятности появления тех или иных лингвистических событий (гл. 5, § 3, п. 1; § 5, п. 4 и др.).

В теории нечетких множеств построен разветвленный набор логических операций, частично повторяющий набор операций классической теории множеств (гл. 1, § 1, п. 3), но включающий и такие специфические операции как концентрация и деконцентрация, сгущение и размывание [26, с. 213]; [65, с. 18—19]. Однако наименее разработанным и наиболее сложным вопросом теории нечетких множеств является количественное определение коэффициента принадлежности μ.

В настоящее время существует три подхода к определению коэффициента μ.

Во-первых, коэффициент принадлежности отождествляют с вероятностью элемента x_i, т. е. $\mu\,(x_i \in A) = P\,(x_i \in A)$; см. гл. 1, § 1, п. 2.

Во-вторых, количественная оценка μ выводится путем сравнения свойств нечеткого множества с характеристиками входящих в него элементов. В своем простейшем виде эту процедуру можно представить себе следующим образом. Предположим, что нечеткое лингвистическое множество A характеризуется некоторым набором равнозначных признаков, а входящий в это нечеткое множество лингвистический элемент x_i имеет m признаков, из которых k являются общими для A и x_i. Тогда степень принадлежности x_i к A равна

$$\mu(x_i \in A) = 1 - \frac{m-k}{m} \quad (0 \leqslant \mu \leqslant 1).$$

Если признаки неэквивалентны, то при расчете следует учитывать их весовые коэффициенты.

В-третьих, величина μ определяется с помощью экспертных оценок, т. е. методом опроса (или голосования) информантов-специалистов. Этот метод, которым мы широко пользовались при измерении информации текста (гл. 5, § 5), дает возможность сгруппировать индивидуальные оценки степени принадлежности x_i лингвистическому множеству A и сообщить этим оценкам количественную меру.

Поскольку при всех этих подходах величина μ нормируется подобно вероятности в интервале от нуля до единицы, к ней могут быть применены многие из тех операций теории вероятностей, математической статистики и теории информации, которые мы применяли в гл. 5—9.

Проиллюстрируем этот подход на следующем примере. Будем рассматривать текст «Курса общей лингвистики» Ф. де Соссюра как нечеткое множество правильных лингвистических положений и подвергнем каждое предложение книги экспертным оценкам с точки зрения степени его принадлежности к указанному множеству. Легко предположить, что такие утверждения Ф. де Соссюра, как «... у речевой деятельности есть и индивидуальная и социальная сторона, причем нельзя понять одну без другой» [59, с. 34] или «... язык необходим, чтобы речь была понятна и производила все свое действие» [59, с. 42], безоговорочно принимаются современными языковедами и вероятно воспринимались как «правильные» теми современниками Ф. де Соссюра, которые разделяли взгляды В. Гумбольдта, Г. Габеленца и И. А. Бодуэна де Куртене. Поэтому величина коэффициента принадлежности этих высказываний к множеству правильных лингвистических высказываний близка к единице ($\mu \simeq 1$).

Вместе с тем в книге Ф. де Соссюра можно найти и такие положения, как «... звук, элемент материальный, не может сам по себе принадлежать к языку» [59, с. 117] или «... в языке нет ничего кроме различий» [59, с. 117]. Эти положения до сих пор оспариваются многими языковедами. Поэтому их принадлежность к рассматриваемому множеству значительно меньше единицы ($\mu \ll 1$). Однако именно эти высказывания вместе с другими не менее парадоксальными утверждениями Ф. де Соссюра составляют эвристическую новизну его «Курса общей лингвистики».

Эвристичность (Eu) тех или иных элементов нечетких лингвистических множеств можно получить, применяя к коэффициенту μ логарифмическую меру подобно тому, как это мы делали, оценивая синтактическую и смысловую информацию, передаваемую отдельными единицами текста (гл. 5, § 5, п. 3 и 4). Тогда получим

$$\mathrm{Eu} = -\log_2 \mu.$$

Если $\mu \simeq 1$, как это имело место для первых двух лингвистических высказываний Ф. де Соссюра, то величина Eu будет близка к нулю. Напротив, для последних утверждений, имеющих $\mu \ll 1$, количественная оценка эвристичности будет достаточно велика (например, если $\mu = 1/16$, то $\mathrm{Eu} = 4$).

Мы воспользовались этим схематичным примером для того, чтобы показать тесную связь и преемственность между традиционными приемами математической лингвистики, описанными в настоящей книге, и вновь создаваемым аппаратом теории нечетких множеств, который, как можно ожидать, даст более глубокое и адекватное описание естественного языка.

ПРИЛОЖЕНИЕ

Таблица I

Значения вероятностей для распределения Пуассона

$$P(X=x)=\frac{\lambda^x}{x!}\,e^{-\lambda}$$

x \ λ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679
1	0905	1637	2222	2681	3033	3293	3476	3595	3659	3679
2	0045	0164	0333	0536	0758	0988	1217	1438	1647	1839
3	0002	0011	0033	0072	0126	0198	0284	0383	0494	0613
4	0000	0001	0003	0007	0016	0030	0050	0077	0111	0153
5	0000	0000	0000	0001	0002	0004	0007	0012	0020	0031
6	0000	0000	0000	0000	0000	0000	0001	0002	0003	0005
7	0000	0000	0000	0000	0000	0000	0000	0000	0000	0001

x \ λ	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
0	0,3329	0,3012	0,2725	0,2466	0,2231	0,2019	0,1827	0,1653	0,1496	0,1353
1	3662	3614	3543	3452	3347	3230	3106	2975	2842	2707
2	2014	2169	2303	2417	2510	2584	2640	2678	2700	2707
3	0738	0867	0998	1128	1255	1378	1496	1607	1710	1804
4	0203	0260	0324	0395	0471	0551	0636	0723	0812	0902
5	0045	0062	0084	0111	0141	0176	0216	0260	0309	0361
6	0008	0012	0018	0026	0035	0047	0061	0078	0096	0120
7	0001	0002	0003	0005	0008	0011	0015	0020	0027	0034
8	0000	0000	0001	0001	0001	0002	0003	0005	0006	0009
9	0000	0000	0000	0000	0000	0000	0001	0001	0001	0002

x \ λ	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	3,0
0	0,1225	0,1108	0,1003	0,0907	0,0821	0,0743	0,0672	0,0608	0,0550	0,0498
1	2572	2438	2306	2177	2052	1931	1815	1703	1596	1494
2	2700	2681	2652	2613	2565	2510	2450	2384	2314	2240
3	1890	1966	2033	2090	2138	2176	2205	2225	2237	2240
4	0992	1082	1169	1254	1336	1414	1488	1557	1622	1680
5	0417	0476	0538	0602	0668	0735	0804	0872	0940	1008
6	0146	0174	0206	0241	0278	0319	0362	0407	0455	0504
7	0044	0055	0068	0083	0099	0118	0139	0163	0188	0216
8	0011	0015	0019	0025	0031	0038	0047	0057	0068	0081
9	0003	0004	0005	0007	0009	0011	0014	0018	0022	0027
10	0001	0001	0001	0001	0002	0003	0004	0005	0006	0008
11	0000	0000	0000	0000	0000	0001	0001	0001	0002	0002
12	0000	0000	0000	0000	0000	0000	0000	0000	0000	0001

Продолжение табл. I

x \ λ	3,5	4,0	4,5	5,0	6,0	7,0	8,0	9,0	10,0
0	0,0302	0,0183	0,0111	0,0067	0,0025	0,0009	0,0003	0,0001	0,0000
1	1507	0735	0500	0337	0149	0064	0027	0011	0005
2	1850	1465	1125	0842	0446	0223	0107	0050	0023
3	2158	1954	1687	1404	0892	0521	0286	0150	0076
4	1888	1954	1898	1755	1339	0912	0573	0337	0189
5	1322	1563	1708	1755	1606	1277	0916	0607	0378
6	0771	1042	1281	1462	1606	1490	1221	0911	0631
7	0385	0595	0824	1044	1377	1490	1396	1171	0901
8	0169	0298	0463	0653	1033	1304	1396	1318	1126
9	0066	0132	0232	0363	0688	1014	1241	1318	1251
10	0023	0053	0104	0181	0413	0710	0993	1186	1251
11	0007	0019	0043	0082	0225	0452	0722	0970	1137
12	0002	0006	0016	0034	0113	0264	0481	0728	0948
13	0001	0002	0006	0013	0052	0142	0296	0504	0729
14	0000	0001	0002	0005	0022	0071	0169	0324	0521
15	0000	0000	0001	0002	0009	0033	0090	0194	0347
16	0000	0000	0000	0000	0003	0014	0045	0109	0217
17	0000	0000	0000	0000	0001	0006	0021	0058	0128
18	0000	0000	0000	0000	0000	0002	0009	0029	0071
19	0000	0000	0000	0000	0000	0001	0004	0014	0037
20	0000	0000	0000	0000	0000	0000	0002	0006	0019
21	0000	0000	0000	0000	0000	0000	0001	0003	0009
22	0000	0000	0000	0000	0000	0000	0000	0001	0004
23	0000	0000	0000	0000	0000	0000	0000	0000	0002
24	0000	0000	0000	0000	0000	0000	0000	0000	0001

Таблица II

Значения функции

$$\varphi(z)=\frac{1}{\sqrt{2\pi}}\,e^{-z^2/2}$$

z	$\varphi(z)$									
	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	0,3989	0,3989	0,3988	0,3986	0,3984	0,3982	0,3980	0,3977	0,3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3272	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0395	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001
4,0	0001	0001	0001	0001	0001	0001	0001	0001	0001	0001
4,5	0000160									
5,0	0000015									

Таблица III

Значения функции Лапласа

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-z^2/2}\,dz$$

x	$\Phi(x)$									
	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0,2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0,3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0,4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0,5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0,6	2257	2291	2324	2357	2389	2422	2454	2486	2517	2549
0,7	2580	2611	2642	2673	2703	2734	2764	2794	2823	2852
0,8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0,9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1,0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1,1	3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1,2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1,3	4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1,4	4192	4207	4222	4236	4252	4265	4279	4292	4306	4319
1,5	4332	4345	4357	4370	4382	4394	4406	4418	4429	4441
1,6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1,7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1,8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1,9	4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2,0	4772	4778	4783	4788	4793	4798	4803	4808	4812	4817
2,1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4858
2,2	4861	4864	4868	4871	4875	4878	4881	4884	4887	4890
2,3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2,4	4918	4920	4922	4924	4927	4929	4931	4933	4934	4936
2,5	4938	4939	4941	4943	4945	4946	4948	4949	4951	4952
2,6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2,7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2,8	4964	4975	4976	4977	4977	4978	4979	4979	4980	4980
2,9	4981	4981	4982	4983	4984	4984	4985	4985	4986	4986
3,00	4986									
3,20	4993									
3,40	4996									
3,60	4998									
3,80	4999									
4,00	49997									
5,00	49999997									

Таблица IV

Значения вероятностей $P(|t| \geqslant t_p)$ для распределения Стьюдента

	$P(\|t\|)$																					
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	∞
0,0	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,0000
0,1	0,937	0,929	0,927	0,925	0,924	0,924	0,923	0,923	0,923	0,922	0,922	0,922	0,922	0,922	0,922	0,922	0,922	0,921	0,921	0,921	0,9203
0,2	874	860	854	851	849	848	847	846	846	845	845	845	845	844	844	844	844	844	844	844	8415
0,3	814	792	784	779	776	774	773	772	771	770	770	769	769	769	768	768	768	768	767	767	7642
0,4	758	728	716	710	706	703	701	700	698	698	697	696	696	695	695	694	694	694	694	693	6892
0,5	705	667	651	643	638	635	632	631	629	628	627	626	625	625	624	624	623	623	623	623	6171
0,6	656	609	591	581	575	570	567	565	563	562	561	560	559	558	557	557	556	556	556	555	5485
0,7	611	556	534	523	515	510	507	504	502	500	498	497	496	495	495	494	493	493	492	492	4839
0,8	570	508	482	469	460	454	450	447	444	442	441	439	438	437	436	435	435	434	434	433	4237
0,9	533	463	434	419	409	403	398	394	392	389	387	385	384	383	382	381	381	380	379	379	3681
1,0	500	423	391	374	363	356	351	347	343	341	339	337	336	334	333	332	331	331	330	329	3173
1,1	470	386	352	333	321	313	308	303	300	297	295	293	291	290	289	288	287	286	285	284	2713
1,2	442	353	316	296	284	275	269	264	261	258	255	253	252	250	249	248	247	246	245	244	2301
1,3	417	323	284	263	250	241	235	230	226	223	220	218	216	215	213	212	211	210	209	208	1936
1,4	395	296	256	234	220	211	204	199	195	192	189	187	185	183	182	181	180	179	178	177	1615
1,5	374	272	231	208	194	184	177	172	168	165	162	159	158	156	154	153	152	151	150	149	1336
1,6	356	251	208	185	170	161	154	148	144	141	138	136	134	132	130	129	128	127	126	125	1096
1,7	339	231	188	164	150	140	133	128	123	120	117	115	113	111	110	108	107	106	105	105	0891
1,8	323	214	170	146	132	122	115	110	105	102	099	097	095	093	092	091	090	089	088	087	0719
1,9	308	198	154	130	116	106	099	094	090	087	084	082	080	078	077	076	075	074	073	072	0574
2,0	295	184	139	116	102	092	086	081	077	073	071	069	067	065	064	063	062	061	060	059	0455
2,1	283	171	127	104	090	080	074	069	065	062	060	058	056	054	053	052	051	050	049	049	0357
2,2	272	159	115	093	079	070	064	059	055	052	050	048	046	045	044	043	042	041	040	040	0278
2,3	261	148	105	083	070	061	055	050	047	044	042	040	039	037	036	035	034	034	033	032	0214
2,4	251	138	096	074	062	053	047	043	040	037	035	034	032	031	030	029	028	027	027	026	0164
2,5	242	130	088	067	054	047	041	037	034	031	030	028	027	025	024	024	023	022	022	021	0124
2,6	234	122	080	060	048	041	035	032	029	026	025	023	022	021	020	019	019	018	018	017	0093
2,7	226	114	074	054	043	036	031	027	024	022	021	019	018	017	016	016	015	015	014	014	0069
2,8	218	107	068	049	038	031	027	023	021	019	017	016	015	014	013	013	012	012	011	011	0051
2,9	211	101	063	044	034	027	023	020	018	016	014	013	012	012	011	010	010	010	009	009	0037

Таблица V

Значения χ^2 для фиксированных уровней значимости и заданных степеней свободы ν

Число степеней свободы ν	Уровень значимости 0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01
1	0,00016	0,00098	0,0039	0,016	2,71	3,84	5,02	6,64
2	0,020	0,051	0,103	0,211	4,61	5,99	7,38	9,21
3	0,115	0,216	0,352	0,584	6,25	7,82	9,39	11,35
4	0,297	0,484	0,711	1,06	7,78	9,49	11,14	13,28
5	0,554	0,831	1,15	1,61	9,24	10,07	12,83	15,09
6	0,872	1,24	1,64	2,20	10,65	12,59	14,45	16,81
7	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48
8	1,65	2,18	2,73	3,49	13,36	15,51	17,54	20,09
9	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67
10	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21
11	3,05	3,82	4,57	5,58	17,28	19,68	21,92	24,73
12	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22
13	4,11	5,01	5,89	7,04	19,81	22,36	24,77	27,69
14	4,66	5,63	6,57	7,79	21,06	23,69	26,12	29,14
15	5,23	6,26	7,26	8,55	22,31	25,00	27,48	30,58
16	5,81	6,91	7,96	9,31	23,54	26,30	28,84	32,00
17	6,41	7,56	8,67	10,09	24,77	27,59	30,19	34,41
18	7,01	8,23	9,39	10,87	25,99	28,87	31,53	34,81
19	7,63	8,91	10,12	11,65	27,20	30,20	32,85	36,19
20	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57
21	8,90	10,28	11,59	13,24	29,62	32,67	35,48	38,93
22	9,54	10,98	12,34	14,04	30,84	33,92	36,78	42,29
23	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64
24	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,98
25	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31
26	12,20	13,84	15,38	17,29	35,56	38,89	41,92	45,64
27	12,88	14,57	16,15	18,11	36,74	40,14	43,19	46,96
28	13,57	15,31	16,93	18,94	37,92	41,34	44,46	48,28
29	14,27	16,05	17,71	19,77	39,09	42,56	45,72	49,59
30	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89
40	22,16	24,43	26,51	29,05	51,80	55,76	59,34	63,69
49	28,94	31,56	33,93	36,82	62,04	66,34	70,22	74,92
50	29,71	32,36	34,76	37,69	63,17	67,51	71,42	76,15
59	36,70	39,66	42,34	45,58	73,28	77,93	82,18	87,17

Таблица VI

Значения $t_{q;\nu}$ и $z_{p;\nu}$ для фиксированных уровней значимости q и заданного числа степеней свободы ν

ν	p	$t_{q;\nu}$: Двусторонний уровень значимости q: 0,10	0,05	0,02	0,01	ν	$t_{q;\nu}$: Двусторонний уровень значимости q: 0,10	0,05	0,02	0,01	q
	q	0,90	0,95	0,98	0,99		0,90	0,95	0,98	0,99	p
1		6,31	12,70	31,82	63,70	18	1,73	2,10	2,55	2,88	
2		2,92	4,30	6,97	9,92	19	1,73	2,09	2,54	2,86	
3		2,35	3,18	4,54	5,84	20	1,73	2,09	2,53	2,85	
4		2,13	2,78	3,75	4,60	21	1,72	2,08	2,52	2,83	
5		2,01	2,57	3,37	4,03	22	1,72	2,07	2,51	2,82	
6		1,94	2,45	3,14	3,71	23	1,71	2,07	2,50	2,81	
7		1,89	2,36	3,10	3,50	24	1,71	2,06	2,49	2,80	
8		1,86	2,31	2,90	3,36	25	1,71	2,06	2,49	2,79	
9		1,83	2,26	2,82	3,25	26	1,71	2,06	2,48	2,78	
10		1,81	2,23	2,76	3,17	27	1,71	2,05	2,47	2,77	
11		1,80	2,20	2,72	3,11	28	1,70	2,05	2,46	2,76	
12		1,78	2,18	2,68	3,05	29	1,70	2,05	2,46	2,76	
13		1,77	2,16	2,65	3,01	30	1,70	2,04	2,46	2,75	
14		1,76	2,14	2,62	2,98	40	1,68	2,02	2,42	2,70	
15		1,75	2,13	2,60	2,95	60	1,67	2,00	2,39	2,66	
16		1,75	2,12	2,58	2,92	120	1,66	1,98	2,36	2,62	
17		1,74	2,11	2,57	2,90	∞	1,64	1,96	2,33	2,58	z_q (z_p)
		0,05	0,025	0,01	0,005		0,05	0,025	0,01	0,005	
		Односторонний уровень значимости $q/2$					Односторонний уровень значимости $q/2$				

Таблица VII

Значения функции

$$P(\lambda) = 1 - K(\lambda) = P(D \geqslant \lambda) = 1 - \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2\lambda^2}$$

λ	$P(\lambda)$	λ	$P(\lambda)$	λ	$P(\lambda)$	λ	$P(\lambda)$
0,32	0,99995	0,78	0,5770	1,24	0,0924	1,70	0,0062
0,33	99991	0,79	5605	1,25	0879	1,71	0058
0,34	99983	0,80	5441	1,26	0836	1,72	0054
0,35	9997	0,81	5280	1,27	0794	1,73	0050
0,36	9995	0,82	5120	1,28	0755	1,74	0047
0,37	9992	0,83	4962	1,29	0717	1,75	0044
0,38	9987	0,84	4806	1,30	0681	1,76	0041
0,39	9981	0,85	4653	1,31	0646	1,77	0038
0,40	9972	0,86	4503	1,32	0613	1,78	0035

Продолжение табл. VII

λ	$P(\lambda)$	λ	$P(\lambda)$	λ	$P(\lambda)$	λ	$P(\lambda)$
0,41	0,9960	0,87	0,4355	1,33	0,0582	1,79	0,0033
0,42	9945	0,88	4209	1,34	0551	1,80	0031
0,43	9926	0,89	4067	1,35	0522	1,81	0029
0,44	9903	0,90	3927	1,36	0495	1,82	0027
0,45	9874	0,91	3791	1,37	0469	1,83	0025
0,46	9840	0,92	3657	1,38	0444	1,84	0023
0,47	9800	0,93	3527	1,39	0420	1,85	0021
0,48	9753	0,94	3399	1,40	0397	1,86	0020
0,49	9700	0,95	3275	1,41	0375	1,87	0019
0,50	9639	0,96	3154	1,42	0354	1,88	0017
0,51	9572	0,97	3036	1,43	0335	1,89	0016
0,52	9497	0,98	2921	1,44	0316	1,90	0015
0,53	9415	0,99	2809	1,45	0298	1,91	0014
0,54	9325	1,00	2700	1,46	0282	1,92	0013
0,55	9228	1,01	2594	1,47	0266	1,93	0012
0,56	9124	1,02	2492	1,48	0250	1,94	0011
0,57	9013	1,03	2392	1,49	0236	1,95	0010
0,58	8896	1,04	2296	1,50	0222	1,96	0009
0,59	8772	1,05	2202	1,51	0209	1,97	0009
0,60	8643	1,06	2111	1,52	0197	1,98	0008
0,61	8508	1,07	2024	1,53	0185	1,99	0007
0,62	8368	1,08	1939	1,54	0174	2,00	0007
0,63	8222	1,09	1857	1,55	0164	2,01	0006
0,64	8073	1,10	1777	1,56	0154	2,02	0006
0,65	7920	1,11	1700	1,57	0145	2,03	0005
0,66	7764	1,12	1626	1,58	0136	2,04	0005
0,67	7604	1,13	1555	1,59	0127	2,05	0004
0,68	7442	1,14	1486	1,60	0120	2,06	0004
0,69	7278	1,15	1420	1,61	0112	2,07	0004
0,70	7112	1,16	1356	1,62	0105	2,08	0004
0,71	6945	1,17	1294	1,63	0098	2,09	0003
0,72	6777	1,18	1235	1,64	0092	2,10	0003
0,73	6609	1,19	1177	1,65	0086	2,20	0001
0,74	6440	1,20	1122	1,66	0081	2,30	0001
0,75	6202	1,21	1070	1,67	0076	2,40	00002
0,76	6104	1,22	1019	1,68	0071	2,50	0000075
0,77	5936	1,23	0970	1,69	0066	3,00	00000003

Таблица VIII

Границы критической области для критерия знаков

n	Односторонние границы					
	2,5%		1%		0,5%	
5	0	5	0	5	0	5
6	1	5	0	6	0	6
7	1	6	1	6	0	7
8	1	7	1	7	1	7
n	5%		2%		1%	
	Двусторонние границы					

Продолжение табл. VIII

n	Односторонние границы					
	2,5%		1%		0,5%	
9	2	7	1	8	1	8
10	2	8	1	9	1	9
11	2	9	2	9	1	10
12	3	9	2	10	2	10
13	3	10	2	11	2	11
14	3	11	3	11	2	12
15	4	11	3	12	3	12
16	4	12	3	13	3	13
17	5	12	4	13	3	14
18	5	13	4	14	4	14
19	5	14	5	14	4	15
20	6	14	5	15	4	16
21	6	15	5	16	5	16
22	6	16	6	16	5	17
23	7	16	6	17	5	18
24	7	17	6	18	6	18
25	8	17	7	18	6	19
26	8	18	7	19	7	19
27	8	19	8	19	7	20
28	9	19	8	20	7	21
29	9	20	8	21	8	21
30	10	20	9	21	8	22
31	10	21	9	22	8	23
32	10	22	9	23	9	23
33	11	22	10	23	9	24
34	11	23	10	24	10	24
35	12	23	11	24	10	25
36	12	24	11	25	10	26
37	13	24	11	26	11	26
38	13	25	12	26	11	27
39	13	26	12	27	12	27
40	14	26	13	27	12	28
41	14	27	13	28	12	29
42	15	27	14	28	13	29
43	15	28	14	29	13	30
44	16	28	14	30	14	30
45	16	29	15	30	14	31
46	16	30	15	31	14	32
47	17	30	16	31	15	32
48	17	31	16	32	15	33
49	18	31	16	33	16	33
50	18	32	17	33	16	34
51	19	32	17	34	16	35
52	19	33	18	34	17	35
53	19	34	18	35	17	36
54	20	34	19	35	18	36
n	5%		2%		1%	
	Двусторонние границы					

Продолжение табл. VIII

n	Односторонние границы					
	2,5%		1%		0,5%	
55	20	35	19	36	18	37
56	21	35	19	37	18	38
57	21	36	20	37	19	38
58	22	36	20	38	19	39
59	22	37	21	38	20	39
60	22	38	21	39	20	40
61	23	38	21	40	21	40
62	23	39	22	40	21	41
63	24	39	22	41	21	42
64	24	40	23	41	22	42
65	25	40	23	42	22	43
66	25	41	24	42	23	43
67	26	41	24	43	23	44
68	26	42	24	44	23	45
69	26	43	25	44	24	45
70	27	43	25	45	24	46
71	27	44	26	45	25	46
72	28	44	26	46	25	47
73	28	45	27	46	26	47
74	29	45	27	47	26	48
75	29	46	27	48	26	49
76	29	47	28	48	27	49
77	30	47	28	49	27	50
78	30	48	29	49	28	50
79	31	48	29	50	28	51
80	31	49	30	50	29	51
81	32	49	30	51	29	52
82	32	50	31	51	29	53
83	33	50	31	52	30	53
84	33	51	31	53	30	54
85	33	52	32	53	31	54
86	34	52	32	54	31	55
87	34	53	33	54	32	55
88	35	53	33	55	32	56
89	35	54	34	55	32	57
90	36	54	34	56	33	57
91	36	55	34	57	33	58
92	37	55	35	57	34	58
93	37	56	35	58	34	59
94	38	56	36	58	35	59
95	38	57	36	59	35	60
96	38	58	37	59	35	61
97	39	58	37	60	36	61
98	39	59	38	60	36	62
99	40	59	38	61	37	62
100	40	60	38	62	37	63
n	5%		2%		1%	
	Двусторонние границы					

Таблица IX

Значения вероятностей для критерия χ^2 Пирсона

χ^2 \ ν	$P(\chi^2)$									
	1	2	3	4	5	6	7	8	9	10
1	0,3173	0,6065	0,8013	0,9098	0,9626	0,9856	0,9948	0,9982	0,9994	0,9998
2	1574	3679	5724	7358	8491	9197	9598	9810	9915	9963
3	0833	2231	3916	5578	7000	8088	8850	9344	9643	9814
4	0455	1358	2615	4060	5494	6767	7798	8571	9114	9473
5	0254	0821	1718	2873	4159	5438	6600	7576	8843	8912
6	0143	0498	1116	1991	3062	4232	5398	6472	7399	8153
7	0081	0302	0719	1359	2206	3208	4289	5366	6371	7254
8	0047	0183	0460	0916	1562	2381	3326	4335	5341	6288
9	0027	0111	0293	0611	1091	1736	2527	3423	4373	5321
10	0016	0067	0186	0404	0752	1247	1886	2650	3505	4405
11	0009	0041	0117	0266	0514	0884	1386	2017	2757	3575
12	0005	0025	0074	0174	0348	0620	1006	1512	2133	2851
13	0003	0015	0046	0113	0234	0430	0721	1119	1626	2237
14	0002	0009	0029	0073	0156	0296	0512	0818	1223	1730
15	0001	0006	0018	0047	0104	0203	0360	0591	0909	1321
16	0001	0003	0011	0030	0068	0138	0251	0424	0669	0996
17	0000	0002	0007	0019	0045	0093	0174	0301	0487	0744
18	0000	0001	0004	0012	0029	0062	0120	0212	0352	0550
19	0000	0001	0003	0008	0019	0042	0082	0149	0252	0403
20	0000	0000	0002	0005	0013	0028	0056	0103	0179	0293
21	0000	0000	0001	0003	0008	0018	0038	0071	0120	0211
22	0000	0000	0001	0002	0005	0012	0025	0049	0089	0151
23	0000	0000	0000	0001	0003	0008	0017	0034	0062	0107
24	0000	0000	0000	0001	0002	0005	0011	0023	0043	0076
25	0000	0000	0000	0001	0001	0003	0008	0016	0030	0053
26	0000	0000	0000	0000	0001	0002	0005	0010	0020	0037
27	0000	0000	0000	0000	0001	0001	0003	0007	0014	0026
28	0000	0000	0000	0000	0000	0001	0002	0005	0010	0018
29	0000	0000	0000	0000	0000	0001	0001	0003	0006	0012
30	0000	0000	0000	0000	0000	0000	0001	0002	0004	0009

Продолжение табл IX

χ^2 \ ν	11	12	13	14	15	16	17	18	19	20
1	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
2	9985	0,9994	0,9998	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
3	9997	9955	9979	9991	0,9996	0,9998	0,9999	1,0000	1,0000	1,0000
4	9699	9834	9912	9955	9977	9989	9995	0,9998	0,9999	1,0000
5	9312	9580	9752	9858	9921	9958	9978	9989	9994	0,9997
6	8734	9161	9462	9665	9797	9881	9932	9962	9979	9989
7	7991	8576	9022	9347	9576	9733	9835	9901	9942	9967
8	7133	7851	8436	8893	9238	9489	9665	9786	9867	9919
9	6219	7029	7729	8311	8775	9134	9403	9597	9735	9829
10	5304	6160	6939	7622	8197	8666	9036	9319	9539	9682
11	4433	5289	6108	6860	7526	8095	8566	8944	9238	9462
12	3626	4457	5276	6063	6790	7440	8001	8472	8856	9161
13	2933	3690	4478	5265	6023	6728	7362	7916	8386	8774

Продолжение табл. IX

$P(\chi^2)$

χ^2 \ ν	11	12	13	14	15	16	17	18	19	20
14	2330	3007	3738	4497	5255	5987	6671	7291	7837	8305
15	1825	2414	3074	3782	4514	5216	5955	6620	7226	7764
16	1411	1912	2491	3134	3821	4530	5238	5925	6573	7166
17	1079	1496	1993	2562	3189	3856	4544	5231	5899	6530
18	0816	1157	1575	2068	2627	3239	3888	4557	5224	5874
19	0611	0885	1232	1649	2137	2687	3285	3918	4568	5218
20	0453	0671	0952	1301	1719	2202	2742	3328	3946	4579
21	0334	0504	0729	1016	1368	1785	2263	2794	3368	3971
22	0244	0375	0554	0786	1078	1432	1847	2320	2843	3405
23	0177	0277	0417	0603	0841	1137	1493	1906	2373	2888
24	0127	0203	0311	0458	0651	0895	1194	1550	1962	2424
25	0091	0148	0231	0346	0499	0698	0947	1249	1605	2014
26	0065	0107	0170	0259	0380	0540	0745	0998	1302	1658
27	0046	0077	0124	0193	0287	0415	0581	0790	1047	1353
28	0032	0055	0090	0142	0216	0316	0449	0621	0834	1094
29	0023	0039	0065	0104	0161	0239	0345	0484	0660	0878
30	0016	0028	0047	0076	0119	0180	0263	0374	0518	0699

Продолжение табл. IX

$P(\chi^2)$

χ^2 \ ν	21	22	23	24	25	26	27	28	29
1	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
2	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
3	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
4	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
5	0,9999	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
6	9994	9997	0,9999	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000
7	9981	9990	9995	9997	0,9999	0,9999	1,0000	1,0000	1,0000
8	9951	9972	9984	9991	9995	9997	0,9999	0,9999	1,0000
9	9892	9933	9960	9976	9986	9992	9995	9997	0,9999
10	9789	9863	9913	9945	9967	9980	9988	9993	9996
11	9628	9747	9832	9890	9929	9955	9972	9983	9990
12	9396	9574	9705	9799	9866	9912	9943	9964	9977
13	9086	9332	9520	9661	9765	9840	9892	9929	9954
14	8696	9015	9269	9466	9617	9730	9813	9872	9914
15	8230	8622	8946	9208	9414	9573	9694	9784	9850
16	7696	8159	8553	8881	9148	9362	9529	9658	9755
17	7111	7634	8093	8487	8818	9091	9311	9486	9622
18	6490	7060	7575	8080	8424	8758	9035	9261	9443
19	5851	6453	7012	7520	7971	8364	8700	8981	9213
20	5213	5830	6419	6968	7468	7916	8308	8645	8929
21	4589	5207	5811	6387	6926	7420	7863	8253	8591
22	3995	4599	5203	5793	6357	6887	7374	7813	8202
23	3440	4017	4608	5198	5776	6329	6850	7330	7765
24	2981	3472	4038	4616	5194	5760	6303	6815	7289
25	2472	2971	3503	4058	4624	5190	5745	6278	6782
26	2064	2517	3009	3532	4076	4631	5186	5730	6255
27	1709	2112	2560	3045	3559	4093	4638	5182	5717
28	1402	1757	2158	2600	3079	3585	4110	4644	5179
29	1140	1449	1803	2201	2639	3111	3609	4125	4651
30	0920	1185	1494	1848	2243	2676	3142	3632	4140

ЛИТЕРАТУРА

1. Аванесов Р. И. Очерки русской диалектологии. М., Учпедгиз, 1949.
2. Аванесов Р. И. Фонетика современного русского литературного языка. М., изд-во МГУ, 1956.
3. Алексеев П. М. Статистическая лексикография. Л., изд-во ЛГПИ, 1975.
4. Арапов М. В., Херц М. М. Математические методы в лингвистике. М., «Наука», 1974.
5. Ахманова О. С. Словарь лингвистических терминов. М., «Советская энциклопедия», 1966.
6. Бектаев К. Б., Пиотровский Р. Г. Математические методы в языкознании. Ч. I. Теория вероятностей и моделирование нормы языка. Алма-Ата, изд-во КазГУ, 1973.
7. Бектаев К. Б., Пиотровский Р. Г. Математические методы в языкознании. Ч. II. Математическая статистика и моделирование текста. Алма-Ата, изд-во КазГУ, 1974.
8. Бектаев К. Б. Статистико-информационная типология текста. Автореф. докт. дисс. Л., Лен. отд. Института языкознания АН СССР, 1975.
9. Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. М., «Наука», 1965.
10. Вентцель Е. С. Теория вероятностей. М., «Наука», 1964.
11. Виноградов В. В. Русский язык. М.—Л., Учпедгиз, 1947.
12. Гиляревский Р. С., Гривнин В. С. Определитель языков мира по письменностям. Изд. 3-е, испр. и доп. М., «Наука», 1965.
13. Гладкий А. В., Мельчук И. А. Элементы математической лингвистики. М., «Наука», 1969.
14. Гнеденко Б. В. Курс теории вероятностей. Изд. 3-е, перераб. М., Физматгиз, 1961.
15. Головин Б. Н. Язык и статистика. М., «Просвещение», 1971.
16. Грамматика русского языка. Т. I. Фонетика и морфология. М., изд-во АН СССР, 1952.
17. Граудина Л. К. Развитие нулевой формы родительного множественного у существительных — единиц измерения. Сб. Развитие грамматики и лексики современного русского языка. М., «Наука», 1964.
18. Зиндер Л. Р., Бондарко Л. В., Вербицкая Л. А. Акустическая характеристика различия твердых и мягких согласных в русском языке. Ученые записки ЛГУ (серия филологических наук), вып. 69, № 325, Л., 1964.
19. Колмогоров А. Н. Основные понятия теории вероятностей. Перевод с немецкого. М.—Л., ОНТИ, 1936.
20. Митропольский А. К. Техника статистических вычислений. Изд. 2-е, перераб. и доп. М., «Наука», 1971.
21. Невельский П. Б. Объем памяти и количество информации Сб. Проблемы инженерной психологии. Психология памяти. Вып. 3. Л., 1965.
22. Обратный словарь русского языка. М., «Советская энциклопедия», 1974.
23. Пиотровский Р. Г. Информационные измерения языка. Л., «Наука», 1968.
24. Пиотровский Р. Г. Моделирование фонологических систем и методы их сравнения. М.—Л., «Наука», 1966.
25. Пиотровский Р. Г. Структурные модели территориальных и жанровых разновидностей языка. Сб. Вопросы романского языкознания. Кишинев, «Штиинца», 1963.
26. Пиотровский Р. Г. Текст, машина, человек. Л., «Наука», 1975.

27. Пиотровский Р. Г. Формирование артикля в романских языках. М.—Л., «Наука», 1960.
28. Пискунов Н. С. Дифференциальное и интегральное исчисления. Изд. 11-е, стереотипн. М., «Наука», 1975.
29. Рождественский Ю. В. Слово в свете данных китайского языка. Сб. Морфологическая структура слова в языках различных типов. М.—Л., изд-во АН СССР, 1963.
30. Романовский В. И. Математическая статистика. Кн. 1. Основы теории вероятностей и математической статистики. Кн. 2. Операторные методы математической статистики. Ташкент, изд-во АН Узб. ССР, 1961, 1963.
31. Сб. Математические методы в языкознании. Рига, «Зинатне», 1969.
32 а, б, в. Сб. Статистика речи и автоматический анализ текста. Л., «Наука», 1971, 1972, 1974.
33. Сб. Статистика текста. Т. I. Лингвостатистические исследования. Минск, изд-во БГУ, 1969.
34. Сб. Статистичні параметри стилів. Київ, «Наукова думка», 1967.
35. Словарь современного русского литературного языка. I—XVII, М.—Л., «Наука», 1950—1965.
36. Смирнов Н. В., Дунин-Барковский И. В. Курс теории вероятностей и математической статистики для технических приложений. Изд. 3-е, стереотипн. М., «Наука», 1969.
37. Тронский И. М Очерки по истории латинского языка. М.—Л., изд-во АН СССР, 1953.
38. Трахтенброт Б. А. Алгоритмы и вычислительные автоматы. М., «Советское радио», 1974.
39. Штейнфельдт Э. А. Частотный словарь современного русского литературного языка. 2500 наиболее употребительных слов. Таллин, изд-во НИИ педагогики ЭССР, 1963.
40. Щерба Л. В. Языковая система и речевая деятельность. Л., «Наука», 1974
41. Яглом А. М., Яглом И. М. Вероятность и информация. Изд. 3-е, перераб. и доп. М., «Наука», 1973.
42. Arley N., Buch K. R. Introduction to the Theory of Probability and Statistics. New-York, 1950. Цит. по русск. пер.: Арлей Н. и Бух К. Р. Введение в теорию вероятностей и математическую стастистику. М., ИЛ, 1951.
43. Bar-Hillel, Y. Some Recent Results in Theoretical Linguistics. Logic, Methodology and Philosophy of Science. Proceedings of the 1960 International Congress. Stanford University Press, Stanford, California, 1962. Цит. по русск. пер.: Бар-Хиллел, И. Некоторые результаты в теоретической лингвистике. Сб. Математическая логика и ее применения. М., «Мир», 1965.
44. Berlin B., Kay P. Basic Color Terms. Their Universality and Evolution. University of California Press, Berkeley and Los Angeles, 1969.
45. Chomsky N., Miller G. A. Introduction to the Formal Analysis of Natural Languages. «Handbook of Psychology», vol. 2, New York, 1963. Цит. по русск. пер.: Хомский Н. и Миллер Дж. Введение в формальный анализ естественных языков. «Кибернетический сборник». Новая серия, вып. 1. М., «Мир», 1965.
46. Church A. Introduction to Mathematical Logic. «Princeton University Press», Princeton, New Jersey, 1956. Цит. по русск. пер.: Черч А. Введение в математическую логику. М., ИЛ, 1960.
47. Coseriu E. Sistema, norma e «parola». «Studi linguistici in onore di Vittore Pisani.» Vol. I, Brescia, Padeia editrice, 1969.
48. Fucks W. Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen. Arbeitsgemeinschaft für Forschung des Landes Nordrhein — Westfalen. Heft 34a. Köln und Opladen, Westdeutscher Verlag, 1955.
49. Greenberg J. H. A Quantitative Approach to the Morphological Typology of Language. «International Journal of American Linguistics». Vol. XXVI, 3, 1960. Цит. по русск. пер.: Гринберг Дж. Квантитативный подход к морфологической типологии языков. Сб. Новое в лингвистике. Вып. 3, М., ИЛ, 1963.

50. Guțu-Romalo V. Stabilirea datei de separație a aromînei de dacoromînă cu ajutorul glotocronologiei. Studii și cercetări lingvistice, X, 4, 1959.
51. Herdan G. Quantitative Linguistics, London, Butterworths, 1964.
52. Hjelmslev L. Prolegomena to a Theory of Language. Supplement to International Journal of American Linguistics. Vol. 19, Indiana University Publications in Antropology and Linguistics. Indianapolis, 1953. Цит. по русск. пер.: Ельмслев Л. Пролегомены к теории языка. Сб. Новое в лингвистике. Вып. 1, М., ИЛ, 1960.
53. Ingarden R. S., Urbanik K. Information without Probability. Colloquium mathematicum. Vol. IX, 1, 1962.
54. Liiv G. Acoustical Features of Estonian Vowels pronounced in Isolation and in three Phonological Degrees of Length. «Eesti NSV Teaduste Akadeemia Toimetised». XI Köide. Ühiskonnateaduste seeria, nr. I, 1962.
55. Mandelbrot B. On the Theory of Word Frequencies and on Related Markovian Models of Discourse. «Structure of Language and its Mathematical Aspects». Proceedings of Symposia in Applied Mathematics. Vol. XII, Providence, Rhode Island, 1961.
56. Marcus S., Nicolau Ed., Stati S. Introducere în lingvistica matematică, București, Editura științifică, 1966.
57. Meillet A. Introduction à l'étude comparative des langues indoeuropéennes. VII édition refondue., P., 1934. Цит. по русск. пер: Мейе А. Сравнительное изучение индоевропейских языков. М.—Л., Соцэкгиз, 1938.
58. Prokosh E. A Comparative Germanic Grammar. Philadelphia, 1939. Цит. по русск. пер.: Прокош Э. Сравнительная грамматика германских языков. М., ИЛ, 1954.
59. Saussure F. de. Cours de linguistique générale. Lausanne — Paris, 1916. Цит. по русск. пер.: Соссюр Ф. де. Курс общей лингвистики. М., Соцэкгиз, 1933.
60. Vaillant A. Le suffixe russe — jaga, «Revue des études slaves», XVI, 1—2, 1938.
61. Van der Warden B. L. Mathematische Statistik. Berlin—Göttingen—Heidelberg, Springer Verlag, 1957. Цит. по русск. пер.: Ван дер Варден Б. Л. Математическая статистика. М., ИЛ, 1960.
62. Vendryes J. Le langage. Introduction linguistique à l'histoire. Paris, 1935. Цит. по русск. пер.: Вандриес Ж. Язык. Лингвистическое введение в историю. М., Соцэкгиз, 1937.
63. Wilks S. S. Mathematical Statistics. New York — London, 1962. Цит. по русск. пер.: Уилкс С. Математическая статистика. М., «Наука», 1967.
64. Yule G. V. and Kendall M. G. An Introduction to the Theory of Statistics, 14-th ed. London, 1950. Цит. по русск. пер.: Юл Дж. Э. и Кендэл М. Дж. Теория статистики. 14-е изд., перераб. и доп. М., Госстатиздат ЦСУ СССР, 1960.
65. Zadeh L. Outline of a New Approach to the Analysis of Complex Systems and Decision Processes. IEEE TSMS. Vol. SMC—3. 1973. Цит. по русск. пер.: Заде Л. А. Основы нового подхода к анализу сложных систем и процессов принятия решений. Математика сегодня (серия математика, кибернетика). М., «Знание», 1974.