Б. Н. ГОЛОВИН

# ЯЗЫК И СТАТИСТИКА

Б. Н. Головин

# ЯЗЫК И СТАТИСТИКА

ИЗДАТЕЛЬСТВО «ПРОСВЕЩЕНИЕ»
Москва 1971

**Головин Б. Н.**

Г61 Язык и статистика. М., «Просвещение», 1970.

190 с.

Автор знакомит студентов-филологов с применением статистики в языкознании, показывает основания и условия вероятностно-статистического изучения языка и речи. В работе приведены статистические таблицы, которые помогут читателям в изучении и применении методики статистического анализа.

7-1-2

БЗ № 15—1971—№ 14



## ОТ АВТОРА

Эта работа — результат десятилетнего знакомства со статистической методикой, десятилетнего опыта ее преподавания студентам-филологам и ее применения в изучении языка и речи.

Теперь известны несколько вариантов статистической методики в языкознании. В книге речь идет об одном из этих вариантов, более близком автору и проверяемом в совместной работе со студентами и аспирантами в процессе учебных и исследовательских занятий.

Можно думать, что преподаватели-филологи высших учебных заведений нуждаются сейчас не в математизации своего языка и своих научных представлений, а в доброй помощи математической науки. Такую помощь филолог почувствует тогда, когда научится применять некоторые, наиболее доступные инструменты математической статистики. Технология изготовления этих инструментов — дело математика, а их применение филологами может быть успешным и при минимальной математической подготовке, *при помощи указаний и разъяснений специальной литературы.*

В вузовских лингвистических курсах «Введение в языкознание» и «Общее языкознание» предполагается изложение сведений о методах и методиках лингвистического анализа. Однако в существующих пособиях по этим курсам соответствующие сведения или крайне скупы, или вообще отсутствуют; и если преподаватель и студент в пособии по введению в языкознание и общему языкознанию все же встречает кое-какие сведения о сравнительно-исторической, сопоставительной и некоторых структурных методиках, то о методиках статистических в этих пособиях, в сущности, нет никаких положительных сведений.

Автор хотел помочь своей работой, прежде всего, преподавателю-лингвисту, читающему общие и специальные лингвистические курсы, студенту-филологу, изучающему и применяющему современные методики лингвистического анализа, учителю-словеснику, готовому применить статистику в решении задач, связанных с развитием речи школьника. Автор хотел показать читателю наиболее доступные инструменты математической статистики, необходимые при решении многих проблем, интересующих и лингвиста, и литературоведа, и психолога, и специалиста в области научно-технической информации, и преподавателя высшей и средней школы, связанного с изучением языка и литературы.

Одну из главных своих задач автор видел в том, чтобы дать читателю по возможности доступное для нематематика изложение всех сведений, относящихся к собственно математической стороне описываемой методики.

Конечно, автору удалось рассказать читателям лишь о ч а с т и проблем, связанных с применением статистической методики в изучении языка. Когда возникала возможность выбора из двух или нескольких решений, автор стремился к наиболее доступным и ясным вариантам методических процедур и их применения. Возможно, с точки зрения строго математической, автор допускал в отдельных случаях упрощения и упущения, и извинить его в таких случаях может лишь желание помочь филологу в овладении непривычной для него методикой.

В работу вошло несколько статистических таблиц; некоторые из них специально построены и рассчитаны автором — для облегчения вычислительной работы филологов.

Автор заранее просит извинения у читателей за отдельные промахи, упущения и просчеты лингвиста в нелегкой, но дорогой для него области математического «оснащения» его науки.

В. В. И в а н о в, В. А. М о с к о в и ч и Н. И. Т о л с т о й любезно согласились прочитать эту работу в рукописи и высказали автору свои советы и замечания, позволившие ему внести в публикуемый текст ряд уточнений и добавлений; автор говорит сердечное спасибо своим рецензентам — за их внимание и квалифицированную, доброжелательную помощь.

## ВМЕСТО ВВЕДЕНИЯ

Кому из читателей не известно одно очень интересное явление — какое-то трудноопределимое сходство отрывков текста одного и того же большого писателя, взятых из разных мест его произведения или даже различных произведений? Ведь совершенно очевидно, что это сходство вызвано не содержанием, а построением речи, какими-то устойчивыми ее особенностями, позволяющими говорить о с т и л е а в т о р а. Часто по двум-трем предложениям читатель устанавливает, кем они могли быть написаны. Можно брать совершенно случайно, не обращая внимания на содержание, куски авторского текста из произведений Гоголя, Герцена, Тургенева, Л. Толстого, Чехова, Паустовского, Леонова, Шолохова — и, как правило, во всех таких кусках Гоголь будет оставаться Гоголем, а Шолохов Шолоховым. То же самое можно сказать и о поэтах: несколько строк Блока сообщат нам непонятным образом, кем они могли быть написаны; то же самое можно сказать о строках Маяковского или Есенина.

Внимательный и достаточно опытный читатель по нескольким предложениям узнает, откуда, из какого текста, они взяты — художественного, научного, газетного или иного.

Это все — факты. Они требуют объяснения.

И стоит, очевидно, задуматься над тем, что если мы узнаем Шолохова, Паустовского или Леонова по нескольким предложениям, значит, есть что-то очень устойчиво-своеобразное в структуре речи этих писателей, сохраняющееся на протяжении большого текста или даже ряда текстов — как будто даже независимо от быстротекущего и изменчивого содержания.

Для того чтобы помочь самим себе справиться с возникшей задачей, представим на минуту поток людей на

большой улице большого города — зимой и весной. Поток движется, непрерывно меняется его облик, и в то же время, вопреки, видимо, этому непрерывному изменению, зимняя толпа остается зимней, а весенняя — весенней. Конечно, и зимой на улице большого города можно встретить людей в демисезонных пальто и шляпах, даже в одних костюмах и с непокрытой головой. Весной же могут встретиться люди в зимних пальто и шапках, еще не заметившие (или не отметившие) наступления весны. Но вторжение в зимний людской поток людей, одетых по-весеннему, или в весенний — одетых по-зимнему — случайно и не меняет общего зимнего или весеннего вида толпы. Дело, таким образом, не в том, что зимой в городской толпе нет людей, одетых по-весеннему, а весной — людей, одетых по-зимнему, а в том, *какое* место занимают те и другие, в *каких* соотношениях находятся они зимой и весной. А это общее соотношение остается без видимых колебаний в разное время зимы или отдельного зимнего дня и в разное время дня весеннего. Конечно, если в середине зимы вдруг наступит сильная оттепель или весной резко похолодает, общий вид городской толпы может измениться.

Не получается ли нечто подобное и с текстом, точнее, с той речевой цепью, тем потоком речи, через который и с помощью которого передается содержание текста?

Теперь уже многочисленные опыты, осуществленные разными лицами в разных странах, не оставляют никаких сомнений в том, что на поставленный вопрос надо отвечать положительно.

Вот одна любопытная иллюстрация, полученная недавно в Горьковском университете студенткой-выпускницей Н. Малиновской. Было взято по шести проб авторской художественной речи у десяти советских русских авторов-прозаиков — и в их числе у Симонова и Шолохова. Каждая проба (на языке статистиков — выборка) равнялась 500 знаменательным словам текста, т. е. представляла собой кусок текста, длиной в 500 знаменательных слов. Пробы брались у Симонова из его романов «Дни и ночи» и «Солдатами не рождаются», а у Шолохова из «Тихого Дона», «Поднятой целины» и «Они сражались за Родину». Проверялась гипотеза о том, что читательское впечатление «стиля Шолохова» и «стиля Симонова» связано с какими-то устойчивыми соотношениями в тексте различных языковых элементов, в частности с устойчивыми и разными для этих писателей частотами хотя бы некоторых частей речи. Выборки (пробы) брались случайно из разных мест разных произведений и состояли (каждая) из слов авторской речи, а не речи персонажей.

Получены такие данные:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Имя существительное: | *Симон.* — | 180 | 175 | 164 | 164 | 155 | 184 |
| | *Шолох.* — | 196 | 195 | 219 | 247 | 243 | 196 |
| Имя прилагательное: | *Симон.* — | 59 | 52 | 47 | 54 | 34 | 48 |
| | *Шолох.* — | 74 | 91 | 89 | 74 | 65 | 70 |
| Местоимение: | *Симон.* — | 54 | 52 | 72 | 67 | 113 | 82 |
| | *Шолох.* — | 38 | 50 | 33 | 18 | 35 | 48 |
| Глагол: | *Симон.* — | 113 | 105 | 123 | 114 | 98 | 113 |
| | *Шолох.* — | 56 | 68 | 90 | 69 | 85 | 91 |
| Наречие: | *Симон.* — | 51 | 75 | 69 | 56 | 70 | 43 |
| | *Шолох.* — | 59 | 59 | 29 | 29 | 30 | 45 |
| Причастие: | *Симон.* — | 21 | 9 | 5 | 19 | 17 | 12 |
| | *Шолох.* — | 56 | 68 | 90 | 69 | 85 | 91 |
| Союзы: | *Симон.* — | 62 | 66 | 90 | 80 | 85 | 75 |
| | *Шолох.* — | 54 | 55 | 38 | 35 | 35 | 29 |

Без формул и инструментов математической статистики видно, что у Шолохова имена существительные, имена прилагательные и причастия заметно активнее, чем у Симонова; у Симонова же заметно активнее, чем у Шолохова, местоимения, глаголы, по-видимому, — наречия, а также союзы.

Принято думать, что части речи слишком абстрактны и инертны, чтобы принимать участие в формировании стилей отдельных авторов. Однако цифры говорят об обратном. И, разумеется, неизмеримо возрастает роль частей речи в формировании функционально-языковых стилей — таких, как публицистический, деловой, научный, художественный.

Ведь наши понятия и представления одеты в форму слов, а сами слова реально известны нам в форме той или иной части речи. Речевой поток подобен потоку людей зимой или весной: меняется соотношение одетых по-весеннему или по-зимнему — меняется и общий вид толпы.

Показанные цифры говорят не только о том, что одни части речи активнее у Симонова, другие — у Шолохова. Они (эти цифры) говорят о том, что эта активность проявляется регулярно: в разных местах разных произведений — следовательно, она закономерно характеризует стиль того или иного автора.

Вот некоторые данные, показывающие активность слов в позициях синтаксиса:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Подлежащие: | *Симон.* — | 65 | 75 | 57 | 53 | 61 | 65 |
| | *Шолох.* — | 50 | 53 | 51 | 59 | 58 | 76 |
| Сказуемые: | *Симон.* — | 99 | 99 | 120 | 109 | 82 | 112 |
| | *Шолох.* — | 51 | 65 | 85 | 67 | 83 | 86 |
| Связки: | *Симон.* — | 12 | 4 | 9 | 9 | 12 | 11 |
| | *Шолох.* — | 3 | 7 | 6 | 2 | 2 | 2 |
| Второстепенные члены: | *Симон.* — | 324 | 322 | 316 | 329 | 335 | 312 |
| | *Шолох.* — | 386 | 375 | 358 | 372 | 357 | 336 |
| Зависят влево: | *Симон.* — | 145 | 163 | 147 | 162 | 150 | 141 |
| | *Шолох.* — | 204 | 224 | 183 | 192 | 153 | 184 |
| Зависят дистантно: | *Симон.* — | 99 | 109 | 104 | 117 | 131 | 109 |
| | *Шолох.* — | 154 | 144 | 138 | 155 | 115 | 135 |
| Число простых предлож.: | *Симон.* — | 20 | 18 | 9 | 9 | 5 | 12 |
| | *Шолох.* — | 13 | 13 | 20 | 17 | 27 | 30 |
| Число сложных предлож.: | *Симон.* — | 21 | 24 | 33 | 32 | 19 | 24 |
| | *Шолох.* — | 11 | 13 | 16 | 13 | 20 | 9 |
| Сочинит. связи: | *Симон.* — | 11 | 17 | 18 | 14 | 7 | 10 |
| | *Шолох.* — | 8 | 8 | 3 | 5 | 9 | 5 |
| Подчинит. связи: | *Симон.* — | 16 | 18 | 29 | 24 | 36 | 29 |
| | *Шолох.* — | 9 | 8 | 8 | 2 | 10 | 2 |

Приведенные цифры показывают, что в речи Симонова активнее слова в роли подлежащих, сказуемых и связки, а в речи Шолохова активнее слова в роли второстепенных членов предложения. Хорошо видно, что у Симонова меньше, чем у Шолохова, зависимых слов, стоящих влево от своих грамматических «хозяев», меньше у Симонова и



слов, зависящих от других слов на расстоянии, т. е. оторванных от своих «хозяев» поставленными между зависимым членом и «хозяином» словами.

У Шолохова активнее простые предложения, у Симонова — сложные; у Симонова заметно больше, чем у Шолохова, сочинительных и подчинительных связей между простыми предложениями в составе сложных.

Как и в морфологии, активность одних синтаксических явлений у Симонова, а других у Шолохова имеет регулярный, т. е., по-видимому, закономерный характер.

А если все это так, разве может лингвист (да и литературовед) не заинтересоваться теми возможностями, которые открывает систематическое изучение языкового функционирования и развития при помощи статистики? А разве только те цифры, которые сообщены читателю на основе части материала одной студенческой дипломной работы, не ставят ряд весьма сложных вопросов, требующих творческого решения? Среди них: а) Связаны ли показываемые статистикой особенности функционирования частей речи, предложений и их членов в речи Симонова и Шолохова некоторыми внутренними зависимостями, т. е. носят ли они системный характер? б) Стоят ли за различиями активности частей речи, членов предложений и предложений у Симонова и Шолохова устойчивые различия художественного содержания произведений двух писателей? в) Нужно ли предполагать, что установленные различия в активности изучавшихся языковых явлений связаны с различиями активности явлений, не изучавшихся в опыте, в частности явлений лексических и лексико-семантических? г) Следует ли думать, что и в необследованных кусках текстов Симонова и Шолохова активность изучаемых элементов речи будет такой же, как и в выборках? д) Есть ли в современной литературе другие писатели, близкие по особенностям речевой структуры к Симонову и Шолохову? е) Вливаются ли речевые структуры Симонова и Шолохова в речевые традиции русской литературы XIX — начала XX в.? ж) Зависят ли и как именно, если зависят, частоты употребления частей речи, членов предложения и предложений от изменения содержания, его динамики на протяжении одного произведения? з) Влияет ли отношение писателя к действительности, которую он изображает, на активность различных явлений языка? И т. д.

Начинать, разумеется, приходится с малого: с накопления фактов, с частот различных явлений языка в разных языковых и речевых стилях. Но и для решения этих «простых» задач нужна не просто арифметика — нужно творчество исследователя, опирающееся и на его интуицию, и на статистическую методику. Постепенно лингвисты все яснее понимают, что для статистического изучения языка и речи нужны, помимо филологических знаний и навыков, еще знания и навыки в области математической статистики. А для того чтобы их приобрести, требуется и отказ от предвзятости, и внутренняя решимость, и усилия, и время. Правда, все эти «потери» с лихвой окупятся получаемыми с помощью статистики результатами — новыми, объективными и действительно творческими. Да и усилия, необходимые филологу для овладения элементарными знаниями и навыками в области математической статистики, не так уж велики.

## ОСНОВАНИЯ И УСЛОВИЯ ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКОГО ИЗУЧЕНИЯ ЯЗЫКА И РЕЧИ

Боязнь статистики медленно, но верно преодолевается большинством лингвистов. На нее начинают обращать внимание и некоторые литературоведы. Однако все еще дает о себе знать непроясненность (или неразъясненность) вопроса о том, на какие реальные, в самом языке существующие основания опирается статистическая методика в языкознании и в каких условиях она может надежно работать.

Если статистическая методика в руках лингвиста (и шире — филолога) всего лишь любопытная игрушка, создающая иллюзию точности и строгости наблюдений и оценок некоторых явлений языка, не стóит, очевидно, и копья ломать, защищая ее применение. Если же эта методика необходима, если ее необходимость диктуется свойствами языка, его природой, если без статистики не может быть достигнуто достаточно полное и разностороннее знание языка,— борьба за внедрение этой методики в работу лингвиста и литературоведа становится принципиальной задачей науки, никак не зависящей от настроений и симпатий отдельных ученых. Вот почему крайне важно вдуматься в реальные, природой языка созидаемые основания вероятностно-статистического исследования языка и речи.

Полезно задуматься и о том, в каких общих и необходимых условиях может быть достигнуто успешное применение статистики для решения лингвистических задач.

Опыты языковедческой и литературоведческой наук, накопленные к настоящему времени знания о языке и его стилях позволяют утверждать, что одним из реальных оснований применения статистики в изучении языка и речи нужно признать объективную присущность языку количественных признаков, количественных характеристик. В неявном виде это признается всеми лингвистами; к тому же многие ученые вынуждены, описывая язык, пользоваться такими количественными понятиями, как «часто», «редко», «употребительно», «неупотребительно», «многочисленный», «много раз», «обычно» и т. д. Но так как такие характеристики имеют лишь общий смысл и никак не проверяются, их надежность недостаточна для построения обоснованной языковой теории.

Конечно, не было случайным обращение к количественным оценкам элементов языка в работах таких больших ученых, как И. А. Бодуэн де Куртенэ, А. М. Пешковский, М. Н. Петерсон, Е. Д. Поливанов, В. В. Виноградов и др. Интересно, что еще в 1938 г. В. В. Виноградов писал: «По-видимому, в разных стилях книжной и разговорной речи, а также в разных стилях и жанрах художественной литературы частота употребления разных типов слов различна. Точные изыскания в этой области помогли бы установить структурно-грамматические, а отчасти и семантические различия между стилями. Но, к сожалению, пока еще этот вопрос находится лишь в подготовительной стадии обследования материала». И далее: «Анализ всех грамматических категорий должен уяснить их относительный функциональный вес в разных стилях литературного языка» [1].

Эти высказывания одного из виднейших представителей классического языкознания не оставляют сомнения в том, что необходимость количественных характеристик в науке о языке была осознана не только сторонниками «математической» лингвистики.

---

[1] В. В. Виноградов. Современный русский язык, вып. I, 1938, стр. 155—156.

По мнению Вяч. Вс. Иванова, «некоторые основные количественные характеристики языка, особенно существенные для лингвиста, носят очень простой характер. Таково, в частности, различие между числом слов (от $10^4$ до $10^6$), числом морфем (несколько тысяч), числом слогов (от нескольких сотен до нескольких тысяч) и числом фонем (от 10 до 80). Согласно высказанному выше предположению, эти соотношения связаны с устройством человеческой памяти».

«Простые количественные отношения между словами, слогами и фонемами позволяют дать классификацию языков, которую можно использовать и при изучении их истории. Так, в случае, если слова в языке односложны, средняя длина слова слишком мала для того, чтобы было возможно членение слова на части; поэтому в языках с односложными словами слово не делится на морфемы. Вместе с тем необходимость различения нескольких тысяч морфем (слов) при ограниченном инвентаре слогов делает необходимым различение слогов посредством музыкального ударения. Поэтому языки с односложными словами-морфемами всегда являются языками с музыкальным ударением (вьетнамский, классический китайский, некоторые центральноафриканские и т. п.). Если в истории языка слова в результате обычного сокращения длины слова становятся односложными, то они перестают членяться на морфемы, и в языке развивается музыкальное ударение (так объясняется ударение в тибетском и классическом китайском языках). Другой существенной закономерностью является связь между числом фонем и средней длиной морфемы: так, в абхазском языке, бзибский говор которого имеет рекордное число фонем (81), корневая морфема обычно бывает однофонемной. В гавайском и других полинезийских языках морфема в среднем состоит из двух слогов (4 фонем), односложные морфемы составляют ничтожный процент общего числа морфем; вместе с тем имеются неразложимые морфемы из 20 фонем (11 слогов), например название рыбы *homohomonukunukuariua*. Знание этих зависимостей можно использовать в сравнительно-историческом языкознании: так, если для какого-либо языка (например, пракартвельского) предполагается однофонемность корня, то это с необходимостью требует предположения о развитой системе фонем, превосходящей среднюю норму (40 фонем). Необходимость количественного исследования явно или неосознанно на практике признается всеми лингвистами, хотя они иногда и заявляют обратное» [1].

Интересные доводы Вяч. Вс. Иванова позволяют перейти к указанию на второе реальное основание, требующее применения вероятностно-статистической методики в изучении языка. Этим вторым основанием является внутренняя зависимость, существующая между качественными и количественными характеристиками языковой структуры. Язык с десятью фонемами даст иное качество звукового облика морфем (а значит, и слов), нежели язык с пятьюдесятью фонемами! Это очевидно.

Можно предполагать, что количественные различия на одном, низшем, уровне дают качественные различия на другом, высшем, уровне. Количество фонем отражается на качестве морфем и слов. Количество морфем сказывается на качестве (уже не только звуковом, но и структурно-семантическом) слов. Количественные характеристики на морфологическом уровне дают о себе знать в качестве синтаксических явлений. Поэтому нужно признать глубоко ошибочным взгляд, в соответствии с которым количественное изучение языка не ведет к качественному, ничего не меняет в уже сложившихся качественных характеристиках, не уточняет их.

Ошибочность этого взгляда становится особенно ощутительной, когда лингвист принимается за изучение языковой структуры речи. Разве нужны особые разветвленные доказательства того, что изменение количественных соотношений между одними и теми же элементами языка (например, частями речи) меняет, и подчас очень резко, качество речи? И разве не об этом писал В. В. Виноградов в 1938 г.? Глубокое научное исследование объективных различий между языковыми и речевыми стилями невозможно без широкого привлечения данных статистики, охватывающих многие и разные элементы и участки языковой структуры. Может быть, сами языковые стили — это не что иное, как основные типы функционирования структуры языка, обслуживающие разные стороны жизни и деятельности общества и отличающиеся друг от друга прежде всего вероятностными характеристиками одних и тех

[1] Вяч. Вс. Иванов. Некоторые проблемы современной лингвистики. «Народы Азии и Африки», 1963, № 4, стр. 176—177.

же элементов и участков языковой структуры. Авторские речевые стили, несомненно, во многом (если и не во всем) определяются устойчивыми для каждого автора соотношениями частот разных элементов языка. Теперь это не гипотеза, а утверждение, опирающееся на известные факты. Вспомним хотя бы те немногие статистические данные, которые получены из текстов Симонова и Шолохова.

Третье реальное основание применимости количественного изучения языка в речи нужно видеть в том, что частоты различных элементов языка в речевом потоке подчиняются, по-видимому, тем или иным статистическим законам. Именно поэтому полученные опытным путем данные о частотах и вероятностях частей речи, некоторых типов предложения, формах глагола говорят о колебаниях частоты каждого изучавшегося элемента языка около некоторой средней величины, причем колебания эти, как правило, статистически закономерны. Оправдывается предвидение русского математика А. А. Маркова, который, указав на недостатки методики, примененной Н. А. Морозовым, говорил: «Только значительное расширение поля исследования (подсчет не пяти тысяч, а сотен тысяч знаков) может придать заключениям некоторую степень основательности, если только границы итогов различных писателей окажутся резко отделенными, а не обнаружится другое весьма вероятное обстоятельство, что итоги всех писателей будут колебаться около среднего числа, подчиняясь общим законам языка» [1].

Обнаружилось именно это другое обстоятельство, которое А. А. Марков признавал весьма вероятным: вот полученные из опыта данные о средних частотах частей речи у русских писателей XIX и XX вв. (данные получены из текстовых выборок длиной каждая в 500 знаменательных слов; было взято по 20 выборок из текстов каждого писателя, из авторской речи; места текста, интуитивно определявшиеся как чуждые художественному тексту, в выборки не включались):

а) глагол: Карамзин — *110*, Пушкин — *110*, Лермонтов — *97*, Гоголь — *97*, Герцен — *94*, Гончаров — *98*, Достоевский — *109*, Л. Толстой — *103*, Тургенев — *107*, Чехов — *127*, Куприн — *77*, Бунин — *87*, А. Толстой — *97*, Гладков — *110*;

б) наречие: соответственно — *29*, *29*, *43*, *45*, *38*, *45*, *56*, *38*, *45*, *42*, *43*, *44*, *31*, *42*;

в) союз: соответственно — *55*, *47*, *45*, *44*, *47*, *74*, *76*, *64*, *53*, *85*, *57*, *53*, *50*, *79*.

Некоторые данные по синтаксису:

а) простые самостоятельные предложения: соответственно — *13*, *26*, *11*, *11*, *14*, *11*, *14*, *15*, *11*, *17*, *20*, *21*, *22*, *28*;

б) сложные предложения: соответственно — *23*, *20*, *22*, *19*, *19*, *18*, *24*, *20*, *20*, *23*, *15*, *15*, *18*, *19*;

в) однородные члены: соответственно — *85*, *73*, *68*, *80*, *95*, *112*, *68*, *76*, *73*, *128*, *47*, *81*, *90*, *61*, *51*.

Эти данные имеют предварительное значение. Они могут быть уточнены и несколько измениться. Но общий их характер останется без заметных перемен. Они несомненно говорят о том, что существует некоторая вероятностная закономерность (или несколько таких закономерностей), управляющая частотами каждого элемента языка. Ведь достаточно внимательного взгляда на ряды чисел, чтобы увидеть относительную устойчивость частот в каждом ряду; применение особых инструментов сравнения частот показало бы, что наши ряды содержат и такие частоты, которые говорят о нарушении общей закономерности отдельными писателями. Но в таких случаях может идти речь о нескольких статистических закономерностях, управляющих речевой деятельностью различных писателей и обнаруживаемых в расхождениях наблюдаемых частот одних и тех же явлений языка.

В мире, в котором мы живем, известны законы двух типов — так называемые динамические и так называемые статистические (вероятностные). Дело, конечно, не в терминах, а в существе различий между теми и другими законами. Действие законов первого типа (т. е. динамических) может быть точно предсказано (например, железо тонет в воде; электролампа загорается при пропускании через ее нить электротока определенного напряжения; вода нормального химического состава и при нормальном атмосферном давлении закипает, если достигает температуры в 100 градусов Цельсия, и т. д.). Действие законов второго типа (т. е. статистических) может быть предсказано лишь в известных пределах от — до, так как проявляется в

[1] А. А. Марков. Об одном применении статистического метода. «Известия Академии наук», 1916, 6-я серия, т. X, вып. 4, стр. 242.

постоянном колебании своих результатов около некоторой средней величины. Статистическим законам подчинены в своем развитии и действии (функционировании) такие явления природы и общественной жизни, которые испытывают влияние большого числа причин, не одинаково направленных, взаимодействующих друг с другом и потому не дающих однозначного результата. Так что нельзя динамические законы противопоставлять статистическим, как причинные непричинным. И те и другие причинны. Однако характер и, если можно так сказать, структура причинности в динамических и статистических законах различны. Статистическим законам подчинены, например, такие сложные явления, как взаимодействие элементарных частиц в микроструктуре вещества, работа человеческого мозга, воздействие школы, пропаганды, искусства на людей, развитие психики ребенка, речевая деятельность, построение речи из элементов языка, развитие и функционирование языка и т. д. В настоящее время многие специалисты-лингвисты и специалисты — математики и физики не сомневаются в том, что язык действует и речь образуется в соответствии со статистическими законами. Вот что пишет, например, американский физик Дж. Пирс: «В нормальном английском тексте, например, в том, который посылается телетайпным аппаратом, отдельные буквы встречаются почти с постоянной частотой. В достаточно длинном тексте почти с постоянной частотой встречаются пары букв, сочетания из трех и четырех букв. Слова и пары слов тоже встречаются почти с постоянной частотой. Далее, с помощью случайного математического процесса, который по желанию может выполнить и машина, мы получим последовательность английских слов или букв со статистическими закономерностями, характерными для английского языка» [1].

Язык может рассматриваться как структура, элементы которой и функционируют в речи, и развиваются, подчиняясь тем или иным вероятностно-статистическим законам. Но если это так, то становится понятной о б ъ е к т и в н а я  н е о б х о д и м о с т ь использования статистической методики, потому что именно эта методика приспособлена специально для улавливания действия различных статистических законов; традиционно-признанные методики, применяемые языкознанием, хороши для качественного описания языковых элементов самих по себе, но они совершенно непригодны для установления и познания генетических и функциональных закономерностей, имеющих статистическую природу. Вот почему приобретает особую значительность постоянное внимание лингвистов к проблемам статистической методики в ее многих вариантах. К числу этих проблем относится и проблема у с л о в и й  у с п е ш н о г о  п р и м е н е н и я  с т а т и с т и к и  в  л и н г в и с т и ч е с к и х  и с с л е д о в а н и я х.

Каковы же эти условия? Пока на этот вопрос ответить полно и вполне удовлетворительно нельзя: еще мал опыт применения статистики в языкознании, и почти нет работ, теоретически рассматривающих этот скромный опыт.

Но все же и теперь отдельные условия успешного применения статистики в языкознании можно назвать. И первым из этих условий хочется видеть с о ю з  с т а т и с т и к и  с  т р а д и ц и о н н ы м и  м е т о д и к а м и  к а ч е с т в е н н о г о  а н а л и з а  я з ы к а. Этот союз необходим хотя бы потому, что лингвист не сможет применить статистику, если не в состоянии строго различать фонемы, морфемы, слова, части речи, члены предложения, типы предложений и т. д. Для того чтобы успешно считать, нужно научиться однозначно узнавать и определять считаемые элементы в различных текстах и устной речи. Статистика сама по себе не может обеспечить распознавание качественных характеристик элементарных единиц языка. Но статистика, опираясь на результаты уже осуществленного лингвистами качественного анализа языковых элементов, показывает закономерности их функционирования и развития и дает основу для качественных оценок уже на новом уровне исследования.

В настоящее время бесплодны и наивны споры о том, какими методами — интуитивными (качественными) или количественными — можно обеспечить успешное развитие науки о языке: нужно не противопоставление одних методов (здесь точнее — методик) другим (например, качественных количественным), а продуманный и гибкий союз различных методик, меняющийся в зависимости от особенностей лингвистических задач.

Вторым условием успешного применения статистики в науке о языке представляется более или менее отчетливое п о н и м а н и е  ученым  т и п о в  л и н г в и с т и-

[1] Дж. Пирс. Символы, сигналы, шумы. М., «Мир», 1967, стр. 69.

ческих задач, решаемых на базе статистики, понимание возможностей статистики в разных областях языковой структуры и на разных ступенях исследовательской абстракции от конкретного языкового или речевого материала.

Какие задачи можно и должно решать при помощи статистической методики в области фонетики языка и звуковой организации речи? Есть ли уверенность в том, что статистика даст положительные результаты в изучении лексики и лексической семантики? Как очерчивается круг главных задач статистического изучения морфологии и синтаксиса? Возможно ли применение статистики в исследовании языковых и речевых стилей? Действительно ли наложен запрет на «поверку гармонии алгеброй» или же это всего лишь иллюзия некоторой части филологов? Если это иллюзия, какие типовые задачи можно предположить в области статистического изучения художественной речи? Как статистически подойти к вопросам речевой культуры и возможны ли объективные, статистические оценки таких качеств речи, как богатство, разнообразие, выразительность и т. д.? Каковы углы статистического зрения на проблемы истории языка?

Конечно, понимание ученым возможностей статистической методики, уточнение типов лингвистических задач, решаемых при ее участии,— это все связано с практикой статистического исследования. Но вместе с тем и в начале своего «статистического» пути ученый должен теоретически представить ожидаемые трудности, возможности и результаты. Нельзя браться за применение статистики, если тип исследовательской задачи таков, что не может принять, по своему существу, вероятностно-статистическое решение.

Третье условие успеха в применении статистической методики — знакомство филолога с минимально-необходимыми для этого статистическими инструментами. Овладев необходимым минимумом, филолог в дальнейшем сможет расширять и уточнять свои знания в новой для него области. Но начинать приходится с малого. И надо знать, что же входит в круг минимально-необходимых лингвисту статистических инструментов. Для того чтобы это знать, придется идти за помощью к математике.



## МИНИМАЛЬНО-НЕОБХОДИМЫЕ СТАТИСТИЧЕСКИЕ ИНСТРУМЕНТЫ

Прежде всего, лингвисту необходимо хотя бы общее представление о статистическом законе и вероятности.

О статистическом законе (в отличие от динамического) речь уже шла в этой книге. Здесь можно лишь добавить, что, по-видимому, все сложные и очень сложные системы (структуры) подчиняются в своем функционировании и развитии статистическим законам. Очень часто в действительности то или иное явление изменяется (функционально или генетически) под влиянием многих воздействий (причин) одновременно, причем эти многие воздействия меняют в некоторых пределах равнодействующую величину совокупного влияния. Но равнодействующая все же определена в границах своих колебаний и подчинена закону.

Простейшие примеры действия статистических законов — подбрасывание игрального кубика или монеты. Хорошо известно, что при достаточно большом числе подбрасываний каждая сторона игрального кубика (а сторон, плоскостей, — шесть) выпадает столько раз (не строго, а приближенно), сколько получится, если разделить общее число подбрасываний на шесть; если подбросим игральный кубик 600 раз, то каждая его сторона выпадет приблизительно по 100 раз, с некоторыми отклонениями от этого идеального случая. Если монету подбросить 500 раз, то каждой своей стороной она выпадет приблизительно 250 раз, но опять-таки с некоторыми отклонениями в ту или другую сторону. Нетрудно понять, что и на игральный кубик, и на монету устойчиво действует одна и та же совокупность причин, влияний — и среди них вес подбрасываемого предмета, его форма, степень однородности его физической структуры, сопротивление воздуха, высота подбрасываний, движение руки человека и т. д. Совокупное влияние многих воздействий, равнодействующая многих причин все время колеблется, но эти колебания случайны и не выходят за некоторые небольшие пределы. Причем, чем больше отклонение от идеального случая, тем реже оно встречается. А это означает, что если сами по себе отклонения возникают случайно, т. е. вследствие не учитываемого для каждого отдельного подбрасывания из-

менения в сочетании многих воздействий, то величина этих отклонений подчинена определенному закону, который и может быть установлен и описан с помощью математики. И именно знание таких законов, управляющих величиной отклонений, позволяет применять статистическую методику как средство сокращения научного эксперимента: по нескольким пробам, выборкам можно судить о той большой совокупности явлений, которая нас интересует и количественные соотношения внутри которой мы хотим определить.

Построив некоторую гипотезу о действии того или иного статистического закона, мы можем, если гипотеза имеет обоснование, говорить о вероятности изучаемого явления (математики говорят — «события»). Понятие «вероятность» не поддается достаточно строгому определению. Поэтому применим «рабочее», нестрогое определение, которое все же поможет нам понять, о чем идет речь. В этом нестрогом смысле вероятность может пониматься как доля изучаемого явления в некотором ряду явлений, ожидаемая на основе гипотезы или предшествующего опыта. Измеряется вероятность отношением числа появлений интересующего нас события в опыте ($m$) к числу всех событий нашего опыта ($n$): $P_a = \frac{m}{n}$.

Когда мы подбрасываем много раз игральный кубик, мы можем заранее, до исхода нашего опыта, сформулировать гипотезу о равной вероятности выпадения кубика каждой из его сторон (плоскостей); такая гипотеза будет отвечать нашему интуитивному представлению о том, что нет никаких видимых причин, которые заставляли бы кубик выпадать одной плоскостью вверх чаще, чем другими.

Между статистическим (вероятностным) законом и вероятностью есть внутренняя зависимость, о которой полезно знать: сама вероятность закономерна, действие изучаемого закона как раз и выражается в сохранении определенной вероятности, изменение вероятности будет говорить и об изменении статистического закона.

И если мы, изучая методами статистики язык и речь, можем каким-либо образом обнаружить вероятность изучаемых фактов и установить, сохраняется или нарушается



эта вероятность, мы тем самым получаем объективное свидетельство действия некоторых законов в функционировании и развитии языка, сохранения и изменения этих законов.

Математическая статистика и дает в руки ученых инструменты наблюдения, с помощью которых можно обнаружить вероятность и установить, сохраняется она или нарушается в определенной области действительности, изучаемой исследователем.

К числу самых элементарных инструментов наблюдения за действием статистических законов, за вероятностью нужно отнести частоту, среднюю частоту и отклонение от средней частоты. Эти термины и соответствующие им понятия входят — наряду с терминами «статистический закон» и «вероятность» — в число наиболее необходимых лингвисту терминов и понятий математической статистики.

Частотой какого-либо явления (факта, «события») называют число его появлений в наблюдаемом отрезке действительности. Этим отрезком может быть любая совокупность считаемых единиц и любая среда, в которой появляются или находятся факты, поддающиеся счету. Понятно, что таким отрезком может быть и текст большего или меньшего объема, большей или меньшей длины.

Например, если мы подбросим игральный кубик 1000 раз и стороной с отметкой «один» он выпадает 170 раз, это число и будет ее частотой. Или если мы возьмем текст длиной в 500 знаменательных слов и насчитаем в нем 100 глаголов, это число мы и назовем наблюдавшейся частотой глагола.

Обычно статистики не считают наблюдаемые и изучаемые факты во всей так называемой «генеральной совокупности» (например, во всех текстах Л. Толстого, если изучается статистически язык Толстого), да это нередко и невозможно. Статистик берет из генеральной совокупности несколько проб, несколько выборок определенного объема и по этим выборкам судит о частотах изучаемых фактов во всей генеральной совокупности. Частоты, показанные отдельными выборками, называются выборочными частотами. Наши примеры (170 выпадений отметки «один» и 100 появлений глагола) — это и есть выборочные частоты одной из сторон игрального кубика и глагола.

Сами по себе выборочные частоты дают очень небольшую информацию о вероятности и статистических законах. Но положение резко меняется, если вводится в действие средняя выборочная частота, или, проще, средняя частота. Есть разные способы и случаи вычисления средних частот. Мы возьмем простейшие и наиболее доступные лингвисту, желающему организовать статистическое изучение текста. Мы берем из текста несколько однородных выборок (однородность определяется интуитивно) одинакового объема (одинаковой длины), например, в 500 или 100 знаменательных слов (или всех слов, считая и служебные). Пусть мы взяли 10 таких выборок. Подсчитываем число наблюдаемых фактов в каждой выборке. Получаем ряд выборочных частот. Чтобы получить среднюю частоту, нам нужно суммировать все выборочные частоты и разделить на число выборок (на число наблюдений). Так, в одном из опытов изучались частоты частей речи в прозе К. Федина. Было взято 10 выборок по 500 знаменательных слов каждая. В выборки включалась только авторская художественная речь (речь персонажей в выборки не вошла, так как явным образом нарушала требование однородности текста). Были получены следующие выборочные частоты имен существительных: 1-я выборка 182, 2-я — 187, 3-я — 218, 4-я — 173, 5-я — 158, 6-я — 201, 7-я — 222, 8-я — 233, 9-я — 213, 10-я — 194; среднюю частоту получим, сложив все выборочные частоты и разделив сумму на 10. Это около 198 существительных в среднем на 500 знаменательных слов.

В том же тексте, в тех же выборках были получены такие частоты имен прилагательных: 69, 71, 83, 60, 43, 73, 72, 59, 69, 71; средняя частота равна приближенно 67 прилагательным на 500 знаменательных слов.

В статистике выборочные частоты принято обозначать буквой $x$ с цифрой-показателем внизу, т. е. $x_1$, $x_2$, $x_3$, $x_4$; обобщенное обозначение любой выборочной частоты данного явления — $x_i$; средняя частота обозначается иксом с чертой, т. е. так: $\bar{x}$.

Роль средних частот в статистическом изучении явлений действительности очень велика. Именно в средних частотах находит своеобразное выражение и отражение та вероятность, которую мы должны знать ради познания статистических законов. Получив средние частоты и обработав их, мы уже можем с известным правом судить о вероятностях. Обработка же средних начинается с того, что наблюдатель вычисляет отклонения выборочных частот от средней частоты; если наблюдавшаяся выборочная частота меньше средней, отклонение получает знак «минус», если выборочная частота больше средней, отклонение получает знак «плюс». Но как ни интересны для наблюдателя-статистика отдельные отклонения сами по себе, он нуждается в некотором их обобщении или усреднении. Такое обобщающее усреднение достигается в статистике обычно двумя путями: а) либо вычисляется среднее абсолютное отклонение, для чего суммируются все отклонения, невзирая на знаки, и сумма отклонения делится на число выборок; б) либо определяется среднее квадратичное отклонение по формуле $\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{\kappa}}$[1]; в формуле $\sigma$ — среднее квадратичное отклонение, $(x_i - \bar{x})$ — отклонения выборочной частоты от средней; $\sum$ — знак суммирования этих отклонений; $\kappa$ — число выборок (наблюдений); если примем $x_i - \bar{x} = a_i$, то формулу можно записать в более простом виде, т. е. $\sigma = \sqrt{\frac{\sum a_i^2}{\kappa}}$. Читается формула так: среднее квадратичное отклонение от средней выборочной частоты равняется корню квадратному из суммы возведенных в квадрат отклонений выборочных частот от их средней, деленной (суммы) на число наблюдений (выборок).

Кстати, два попутных замечания: 1) формула сообщена здесь в своем простейшем виде для случая, когда все выборки равны по длине или объему; 2) величина $\frac{\sum (x_i - \bar{x})^2}{\kappa}$ носит в статистике название дисперсии и обозначается знаком $\sigma^2$.

В математической статистике пользуются обычно не средним абсолютным, а именно средним квадратичным от-

[1] В статистике применяется и другой вид формулы для расчета дисперсии ($\sigma^2$) и среднего квадратичного отклонения ($\sigma$): $\sigma^2 = \frac{\sum x_i^2}{\kappa} - \bar{x}^2$; отсюда $\sigma = \sqrt{\frac{\sum x_i^2}{\kappa} - \bar{x}^2}$.

клонением — из-за его чисто математических преимуществ, рассматривать которые здесь не место.

Допустим, из какого-то текста были взяты 5 выборок по 500 знаменательных слов и были получены следующие частоты глаголов: 1-я выборка — 95, 2-я — 87, 3-я — 94, 4-я — 104, 5-я — 100; нужно определить среднее квадратичное отклонение. Для этого прежде всего вычисляем среднюю частоту: $x = \frac{95+87+94+104+100}{5} = 96$, затем вычисляем отклонения от средней частоты для каждой выборки: 1-я: 95—96=—1; 2-я: 87—96=—9; 3-я: 94—96=—2; 4-я: 104—96= +8; 5-я: 100—96= +4. Теперь можно вычислить и среднее квадратичное отклонение; для этого сначала возведем каждое из отклонений в квадрат и получим числа 1, 81, 4, 64, 16; затем суммируем все квадраты отклонений, получим число 166; разделим 166 на число выборок, т. е. на 5, получим 53,2; извлечем из этого числа квадратный корень, получим 7,29; это и есть величина среднего квадратичного отклонения.

Разумеется, реально, в практике вычислений, вся процедура определения среднего квадратичного отклонения протекает заметно быстрее. Очень полезно использовать таблицу, например, такую:

| Выборки | Выборочные частоты их отклонения от средней частоты и квадраты этих отклонений | | | | | |
|---|---|---|---|---|---|---|
| | Имена | | | Глаголы | | |
| | $x_i$ | $a_i$ | $a_i^2$ | $x_i$ | $a_i$ | $a_i^2$ |
| 1-я | 199 | —3 | 9 | 95 | —1 | 1 |
| 2-я | 205 | +3 | 9 | 87 | —9 | 81 |
| 3-я | 195 | —7 | 49 | 94 | —2 | 4 |
| 4-я | 201 | —1 | 1 | 104 | +8 | 64 |
| 5-я | 210 | +8 | 64 | 100 | +4 | 16 |
| Сумма | 1010 | 0 | 132 | 480 | 0 | 166 |
| $\bar{x}$ | 202 | | | 96 | | |
| $\sigma$ | | 5,14 | | | 7,29 | |



Математическая статистика утверждает, что в практике статистического изучения лучше применять не формулу

$$\sqrt{\frac{\sum a_i^2}{k}},$$

а другую, отличающуюся только тем, что в знаменателе дроби под знаком корня стоит не $k$, а $k$—1. По этой уточненной формуле вычисляется так называемая несмещенная оценка среднего квадратичного отклонения, которая служит своеобразным уточнителем выборочного среднего квадратичного отклонения; потребность же в уточнении возникает потому, что формула квадратичного отклонения имеет в виду некий идеальный случай, теоретические соотношения между средней частотой, отклонениями и средним квадратичным отклонением. Выборочные же величины всегда несколько отличаются от теоретических, они менее точны, менее строги; поэтому и возникает необходимость внести поправку, которую и дает вторая формула. Однако для решения лингвистических задач можно пользоваться и основной формулой, так как большой точности статистическое изучение языка не требует.

Итак, мы знакомы с шестью весьма важными терминами и понятиями математической статистики — вероятностью, статистическим законом, выборочной частотой, средней выборочной частотой, отклонением от средней частоты и средним квадратичным отклонением. Этими понятиями почти исчерпывается круг фундаментально-необходимых лингвисту понятий, заимствуемых из математической статистики. Введем еще лишь одно; остальные же будут вводиться попутно с изложением других вопросов методики, причем эти «остальные» окажутся производимыми от основных.

Это еще одно понятие выражается термином «вероятная ошибка в определении средней частоты». Дело в том, что наши выборочные данные не дают нам знания той действительной средней, которая характеризует всю изучаемую совокупность. Например, если на основании 20 выборок мы получили выборочную среднюю частоту глагола в текстах Пушкина в 110 единиц, это еще не означает, что «действительная средняя» всех текстов Пушкина, из которых брались выборки, однородных этим выборкам по структуре речи, равна также 110. Эту

действительную среднюю мы не знаем. Но именно для того, чтобы иметь о ней приближенное представление, нам было нужно взять наши выборки и определить выборочную среднюю частоту. Действительная средняя должна быть где-то около нашей средней. Но где именно? В каком интервале частот? Для ответа на такие вопросы и используется знание вероятной ошибки в определении средней (как это делается, об этом речь пойдет позже). Эта ошибка находится в известной зависимости как от величины средней, так и от отклонений от нее, а также от количества наблюдений. Нетрудно понять, что, чем устойчивее наши частоты, чем меньше они разбросаны вокруг средней, тем надежнее сама средняя; с другой стороны, чем больше мы сделали проб, чем больше взяли выборок, тем надежнее полученный результат, т. е. величина средней частоты.

Одним словом, вероятная ошибка в определении средней вычисляется по формуле $L=\frac{t\sigma}{\sqrt{\kappa}}$. В этой формуле $L$ — величина ошибки, $t$ — особый коэффициент, зависящий от числа наблюдений (выборок), он берется из таблицы; $\sigma$ — знакомое нам среднее квадратичное отклонение или, вместо него, можно взять $s=\sqrt{\frac{\sum a_i^2}{\kappa-1}}$ (это даже лучше); $\kappa$ — число наблюдений (выборок).

$t$ при пяти выборках нужно брать равным 2,78, при десяти — 2,26, при пятнадцати — 2,15, при двадцати — 2,09, при двадцати пяти — 2,06, при тридцати — 2,05.

Эти коэффициенты обеспечивают 95-процентную надежность показаний формулы. Как это понять?

Пусть один опыт по статистическому изучению некоторого текста $A$ состоит из обработки данных 10 выборок, и эти данные показали ошибку в определении средней, равную 10 единицам, при 95-процентной надежности. Это значит, что полученные нами данные позволяют предположить, что если бы мы текст $A$ обследовали не один раз, а 100, т. е. осуществили бы 100 опытов, аналогичных уже осуществленному, причем выборки во всех опытах были бы аналогичны тем, которые были взяты в первом опыте, то в 95 опытах средние частоты не отличались бы от найденной в первом опыте более чем на 10 единиц в ту или другую сторону, т. е. лежали бы в пределах от 100+10 до



100—10 (т. е. в интервале 90—110); в пяти опытах из ста средняя частота могла бы выйти за эти пределы.

Обычно статистики рекомендуют 95-процентную надежность определения ошибки в исчислении средней частоты. Но если мы можем довольствоваться и 92-процентной надежностью, то в формулу ошибки можно ввести постоянный коэффициент 2 и применять его при любом числе выборок от 10 и более, — в таком случае надежность не будет менее 92%.

Таковы некоторые элементарные сведения о самых необходимых лингвисту понятиях и терминах математической статистики, являющихся одним из условий успешного применения в изучении языка статистической методики.

Пусть, читатель, наши знания в области математической статистики весьма элементарны. Но и они уже позволяют ставить вполне удовлетворительные по результатам статистические опыты. Теперь ничто не мешает нам взять из интересующего нас текста (или текстов), например, по 10 выборок, каждая длиной в 500 (можно, конечно, и в 250, и в 1000 и т. д.) знаменательных слов, а точнее их словоупотреблений. Допустим, что текстов мы взяли всего два и хотим сравнить в них частоты глаголов. Первый текст (назовем его *ТА*) дал частоты: 95, 98, 89, 105, 102, 85, 111, 115, 93, 107; второй текст (назовем его *ТБ*) дал частоты: 98, 112, 114, 108, 106, 122, 95, 87, 125, 133. Найдем средние выборочные частоты в наших двух текстах. Если вычисления будут правильными, текст *ТА* покажет среднюю частоту, равную 100, а текст *ТБ* — частоту, равную 110. Проверьте, пожалуйста, читатель!

Арифметически 110 больше 100. Но статистика имеет свое представление о равенстве и неравенстве. Об этом еще пойдет речь. Но и сейчас мы сможем сравнить наши средние частоты не арифметически, а статистически. Для этого: а) вычислим отклонения от средних частот в текстах *ТА* и *ТБ*; б) возведем каждое отклонение в квадрат; в) вычислим суммы возведенных в квадрат отклонений для текста *ТА* и текста *ТБ*; г) найдем по формуле несмещенной оценки среднего квадратичного отклонения эти несмещенные оценки для текста *ТА* и текста *ТБ*; д) по формуле ошибки наблюдения $L=\frac{2s}{\sqrt{\kappa}}$ (коэффициент 2 возьмем пока для простоты вычислений) определим эти ошибки для текста *ТА* и текста *ТБ*; е) найдем (прибавляя к выборочным средним

ошибку и отнимая ее) границы действительных средних. Они в нашем примере должны быть такими: для *ТА* от 94 до 106, для *ТБ*— от 101 до 119 (результаты вычислений округлены). Проверьте себя, читатель!

## СТАТИСТИЧЕСКАЯ ОЦЕНКА РАСХОЖДЕНИЙ МЕЖДУ ВЫБОРОЧНЫМИ ЧАСТОТАМИ

Если статистическое изучение языка или речи ведется путем выборок из текста и каждая выборка имеет одну и ту же «длину», наблюдатель (лингвист) оказывается перед необходимостью как-то оценить те колебания частот одного и того же языкового явления, которые неизбежно возникают и в которых заключена информация о действии статистических законов и их нарушениях.

Предположим, что десять текстовых выборок по 500 знаменательных слов (словоупотреблений) каждая дали такой ряд частот глагола: 98, 87, 102, 105, 123, 108, 85, 78, 110, 104. О чем говорят наблюдателю эти колебания? Могла ли дать их одна и та же вероятность, один и тот же статистический закон? Если так, то замеченные колебания случайны и, следовательно, статистически закономерны. Или, может быть, наиболее заметные отклонения от средней частоты возникли вследствие нарушения статистического закона, вследствие изменения вероятности на протяжении нашего опыта, и если так, колебания частот не случайны, они существенны, они незакономерны для одной и той же вероятности.

Лингвист заинтересован в том, чтобы каким-то образом установить, случайны или существенны отклонения выборочных частот от их средней. Как же сделать это?

Математическая статистика, в числе многих своих инструментов, с помощью которых решаются различные задачи статистического изучения, имеет инструмент, называемый «хи-квадрат критерий» и обозначаемый греческой буквой $\chi^2$. Вот формула, по которой вычисляется величина «хи-квадрат» в таких случаях, как наш, т. е. когда все выборки имеют одинаковую длину:

$$\chi^2 = \frac{\sum (x_i - \bar{x})^2}{\bar{x}} .$$

В этой формуле $x_i$ — наблюдаемые частоты, $\bar{x}$ — средняя выборочная частота, $\sum$ — знак суммирования. Если обозначить отклонение выборочной частоты от средней буквой «а», как это было сделано ранее, то формула «хи-квадрат» получит менее громоздкий вид: $\chi^2 = \frac{\sum a_i^2}{\bar{x}}$. Эта формула (в первой или второй ее записи, все равно) читается так: «хи-квадрат» равен сумме квадратов отклонений от средней частоты, деленной на среднюю частоту. Иначе говоря «хи-квадрат» — не что иное, как отношение суммы квадратов отклонений от средней частоты к этой частоте.

Математики установили, что для одной и той же вероятности величина этого отношения подчиняется определенному закону «распределения частот», т. е. одно отношение встречается часто, другое — реже, третье еще реже и т.д. Математики составили особые таблицы, в которых указано допустимое теорией отношение «хи-квадрат», допустимая теорией величина, которую и можно использовать для оценки наблюдавшегося в опыте расхождения частот: в числителе нашей формулы как раз и стоят символы, указывающие на отклонение выборочных частот от их средней.

Критерий «хи-квадрат» часто называют критерием согласия. Чего с чем? По-видимому, опытных, вычисленных по формуле величин с величинами теоретическими, соответствующими закону случайного варьирования одной и той же вероятности. Значит, получив из некоторого опыта величину «хи-квадрат», мы должны сличить ее с соответствующей теоретической величиной, для этого и приходится обращаться к особой таблице.

В таблице указаны различные числовые значения «хи-квадратов», соответствующие различным случаям расхождений между средней частотой и отклонениями от нее. Величины «хи-квадрат» соответствуют каждая определенной «степени свободы» (горизонтальная строка) и определенной вероятности (вертикальный столбец). Понятие «степень свободы» оставим пока без пояснений — ввиду его сложности; но примем без доказательств, что при сравнении нескольких выборочных частот — и при условии, что все выборки имеют равную длину,— число степеней свободы будет на единицу меньше числа выборок. Понятие же «вероятность большего значения» может получить некоторые несложные пояснения. Например, в строке таблицы, соответствующей девяти степеням свободы (то, что нам нужно: у нас в опыте было десять выборок), величине «хи-квадрат» 16,92 соответствует показатель веро-

ятности — 0,05 (т. е. если дать другое, процентное выражение, то получится 5%). Значит, такое расхождение частот, которое дает величину «хи-квадрат», равную 16,92, или бо́льшую, встречается в пяти теоретических случаях из ста. Математическая статистика утверждает, что, если выборочное расхождение частот (или выборочное отклонение от некоторой теоретической средней) дает величину «хи-квадрат», не превышающую его («хи-квадрата») теоретического значения, соответствующего 5% вероятности, измеряемые расхождения частот можно признать случайными; если же выборочный «хи-квадрат» превосходит величину теоретического (табличного), соответствующую пятипроцентной вероятности, расхождение частот признается существенным, и выдвинутая гипотеза отвергается. Но почему появилось выражение «выдвинутая гипотеза»? А потому, что, когда при помощи критерия «хи-квадрат» мы сравниваем расхождения (отклонения) частот, мы сознательно или бессознательно проверяем некоторые гипотезы, предположения. Например, если мы вернемся к задаче о частотах глагола, то обнаружим гипотезу, требующую проверки и допускающую такую формулировку: все наблюдаемые частоты суть проявления одной и той же вероятности и потому их отклонения от их средней частоты случайны. Произведя необходимые вычисления, мы получим величину «хи-квадрат» 16,2, т. е. несколько меньше теоретической (табличной). Это позволяет нам принять гипотезу о случайности отклонения частот 98, 87, 102, 105, 123, 108, 85, 78, 110, 104 от их средней, т. е. от 100.

Предложенное далее извлечение из более полной таблицы «хи-квадратов» вполне достаточно для решения многих исследовательских задач на основе выборочной методики при использовании реальных текстов.

Как же пользоваться таблицей? Допустим, что было сделано пять выборок по 500 знаменательных слов каждая. Были получены частоты имен прилагательных: 75, 70, 82, 68 и 80; средняя частота — 75. Мы хотим проверить гипотезу о том, что все пять выборок взяты из совокупности с одной и той же вероятностью имен прилагательных; иначе говоря, это гипотеза о том, что все отклонения частот от их общей средней носят случайный, не существенный характер. Для проверки гипотезы вычисляем величину «хи-квадрат» и находим, что она равна 1,97. В таблице выбираем строку, соответствующую четырем степе-



**Извлечение из таблицы числовых значений $\chi^2$**

| Число степеней свободы | Вероятность большего значения | | | | | |
|---|---|---|---|---|---|---|
| | 0,95 (95%) | 0,75 (75%) | 0,50 (50%) | 0,25 (25%) | 0,10 (10%) | 0,05 (5%) |
| 1 | — | 0,10 | 0,45 | 1,32 | 2,71 | 3,84 |
| 2 | 0,10 | 0,58 | 1,39 | 2,77 | 4,61 | 5,99 |
| 3 | 0,35 | 1,21 | 2,37 | 4,11 | 6,25 | 7,81 |
| 4 | 0,71 | 1,92 | 3,36 | 5,39 | 7,78 | 9,49 |
| 5 | 1,15 | 2,67 | 4,35 | 6,63 | 9,24 | 11,07 |
| 6 | 1,64 | 3,45 | 5,35 | 7,84 | 10,64 | 12,59 |
| 7 | 2,17 | 4,25 | 6,35 | 9,04 | 12,02 | 14,07 |
| 8 | 2,73 | 5,07 | 7,34 | 10,22 | 13,36 | 15,51 |
| 9 | 3,33 | 5,90 | 8,34 | 11,39 | 14,68 | 16,92 |
| 10 | 3,94 | 6,74 | 9,34 | 12,55 | 15,99 | 18,31 |
| 14 | 6,57 | 10,17 | 13,34 | 17,12 | 21,06 | 23,68 |
| 15 | 7,26 | 11,04 | 14,34 | 18,25 | 22,31 | 25,00 |
| 19 | 10,12 | 14,56 | 18,34 | 22,72 | 27,20 | 30,14 |
| 20 | 10,85 | 15,45 | 19,34 | 23,83 | 28,41 | 31,41 |
| 24 | 13,85 | 19,04 | 23,34 | 28,24 | 33,20 | 36,42 |
| 25 | 14,61 | 19,94 | 24,34 | 29,34 | 34,38 | 37,65 |
| 29 | 17,71 | 23,57 | 28,34 | 33,71 | 39,09 | 42,56 |
| 30 | 18,49 | 24,48 | 29,34 | 34,80 | 40,26 | 43,77 |

ням свободы. (Только четыре частоты из пяти могут в формуле «хи-квадрат» принимать любые значения, если дана еще и средняя частота; пятая частота не свободна, она задана четырьмя свободными частотами и общей для всех пяти частот средней; поэтому математик скажет, что мы имеем четыре степени свободы.) Мы увидим, что полученный нами «хи-квадрат» соответствует примерно 70% вероятности большего значения. Хорошо это или плохо? Принимается или отвергается наша гипотеза?

Статистики применяют понятие «границы существенности»; ими являются те критические вероятности, переход через которые явно свидетельствует о существенных колебаниях или расхождениях частот. Таких границ две; ими обычно считаются 95-процентная и 5-процентная вероятность большего значения. Почему? Потому, что 95-процентная вероятность слишком велика: соответствующая ей или меньшая величина «хи-квадрат» встречается всего

пять раз на сто; мы осуществили не 100 опытов, а всего один — и сразу получили столь редкую величину. Так как это событие маловероятно, гипотезу о случайности колебания интересующих нас частот мы должны были бы отвергнуть; мы должны были бы выдвинуть иную гипотезу — о нестатистическом, слишком жестком характере зависимости наших частот от каких-то условий. 5-процентная вероятность большего значения принимается как вторая граница существенности потому, что она слишком мала: соответствующая ей или большая величина «хи-квадрат» встречается также всего лишь 5 раз на сто. И если именно такую величину мы встретили в первом же опыте, она не заслуживает доверия, и мы должны отказаться от нашей гипотезы и заменить ее иной — об отсутствии статистической закономерности, о колебании вероятности в пределах той совокупности фактов, соотношения частот внутри которой нас интересуют.

Но вернемся к величине «хи-квадрат». Она была в нашем опыте равна 1,97 при четырех степенях свободы. Эта величина соответствует примерно (посмотрим нашу таблицу!) 70% вероятности большего значения; это довольно далеко от верхней границы существенности, и мы принимаем сформулированную ранее гипотезу о случайности колебания наблюдавшихся частот около их общей средней частоты; нет основания отвергать предположение о действии в той совокупности, из которой были взяты выборки, одной и той же вероятности.

Итак, мы применяем таблицу «хи-квадратов» для сравнения выборочной величины $\chi^2$ с теоретической и используем данные сравнения для проверки согласования некоторой гипотезы с реальностью, показанной колебаниями частот нескольких выборок из одной и той же совокупности фактов.

Нужно иметь в виду, что ни один критерий из предлагаемых математической статистикой не дает вполне определенного ответа на вопрос: «Верна или не верна данная статистическая гипотеза?» Ответы на подобные вопросы имеют вероятностный характер. Это нужно помнить и не требовать от математической статистики невозможного.

И критерий «хи-квадрат», даже очень осмотрительно примененный, не позволяет исследователю делать категорические суждения о вероятностях, частотах, их колебаниях и т. д. Все эти суждения сами должны иметь вероятностный характер, т. е. они имеют всегда некоторый допуск на возможную ошибку. Но сама возможность ошибки, величина этой возможности, обычно применяемым критерием (в частности, критерием «хи-квадрат») измеряется. Все это строго соответствует характеру изучаемых при помощи статистики вероятностных законов.

Математическая теория и многочисленные опыты применения критерия «хи-квадрат» говорят о том, что он удобен для решения многих задач (определение величины и характера колебания частот около средней, сравнение обобщенных представлений о величине колеблемости изучаемых явлений, статистическое сравнение частот одного и того же явления в двух разных совокупностях и т. д.). Вместе с тем очевидно, что «хи-квадрат» обладает такой большой точностью, такой большой бракующей гипотезы силой, что его применение может привести к отказу от гипотезы, когда это можно было бы и не делать. Проще говоря, «хи-квадрат» часто дает отрицательный ответ на вопрос «Случайны ли расхождения этих двух частот?» — в то время когда по существу гипотезу о несущественности, случайности их расхождения можно было бы принять. И чем больше сравниваемые частоты, тем сильнее чувствительность и бракующая сила критерия «хи-квадрат». При очень малых частотах «хи-квадрат» оказывается слишком либеральным судьей и иногда может пропустить гипотезу о случайности расхождения частот — в то время как ее нужно было бы из осторожности отвергнуть.

Мне кажется, что лингвистам можно рекомендовать применение критерия «хи-квадрат», когда сравниваемые частоты находятся в промежутке от десятка-двух до нескольких сотен; этот промежуток намечен очень условно и его указание никак не означает, что за его пределами частоты не могут оцениваться с помощью «хи-квадрата».

Вот несколько примеров статистических задач, которые может решать лингвист с помощью критерия «хи-квадрат»:

а) из текста взяты равные по объему выборки, давшие ряд частот. Можно ли думать, что колебания частот случайны, т. е. объясняются лишь законами статистического варьирования одной и той же средней? Решение аналогичной задачи было показано. Обследование ряда текстов под этим углом зрения может дать объективную информацию о колебаниях изучаемых языковых явлений, зависимости этих колебаний от различного рода условий, в которых

оказываются изучаемые языковые факты, и т. д. Нужно, по-видимому, признать, что колеблемость языковых элементов может использоваться как объективная характеристика и самих элементов и того текста, в котором полученная колеблемость возникла. Величиной колеблемости (а она может оцениваться величиной «хи-квадрат») характеризуется устойчивость или неустойчивость различных элементов языковой структуры в разных условиях их текстового применения;

б) в опыте получены две частоты одного и того же явления языка в двух текстовых совокупностях, выборки из которых были равного объема (выборки, разумеется, могут «отмеряться» не только количеством знаменательных слов, но иными способами, например количеством страниц, числом строк и т. д., если страницы и строки примерно одинаковы по размеру, т. е. по числу строк в странице и по числу знаков в строке). Возникает задача статистически сравнить частоты, т. е. ответить на вопрос: «Существенны или случайны расхождения полученных в опыте частот?» Найдем общую выборочную среднюю частоту, т. е. сумми-

[redacted]

чину каждой выборки: а) 75+100=175; б) 530+970=1500; в) 175 : 1500=0,116; г) 0,116×530=61,5; д) 0,116×970=113,5. Теперь можно пустить в ход формулу:

$$\chi^2 = \frac{(75-61{,}5)^2}{61{,}5} + \frac{(100-113{,}5)^2}{113{,}5} = 4{,}57.$$

При одной степени свободы «критическая» величина $\chi^2$, по данным таблицы, равна 3,84; из опыта мы получили величину 4,57, т. е. большую, чем критическая; поэтому мы не можем сохранить гипотезу о несущественности, случайности того расхождения частот, которое было обнаружено в двух не равных по длине выборках.

Итак, были рассмотрены два типа задач, в решении которых целесообразно применять критерий «хи-квадрат»: во-первых, обобщенная оценка величины и характера колеблемости частот в их ряду и, во-вторых, оценка величины расхождения двух частот.

Первая задача может решаться при помощи менее точного, но и более доступного инструмента, названного коэффициентом вариации, а вторая может заменяться сход-

Вернемся к задаче, которая решалась нами в связи с введением понятия о «хи-квадрате». Получен ряд частот (из выборок равного объема): 98, 87, 102, 105, 123, 108, 85, 78, 110, 104 при средней частоте — 100. Каков коэффициент вариации? Вычислив среднее квадратичное отклонение, получим — 12,7. Отсюда коэффициент вариации равен 12,7 : 100 : 100=12,7%. Такой коэффициент вариации признается вполне допустимым для гипотезы о случайности варьирования частот. Принято величину коэффициента вариации считать «большой», т. е. вызывающей недоверие к гипотезе о случайном варьировании частот, тогда, когда он превосходит 40%. Конечно, это лишь весьма приближенно взятая граница существенности, но она вполне оправдывает себя в тех случаях, когда не требуется большая точность. Если же нужна именно большая точность, лучше отказаться от коэффициента вариации и применить критерий «хи-квадрат». Показанное ранее применение этого критерия к задаче, только что предложенной вторично, позволило сохранить гипотезу о случайности колебания частот около средней; то же самое нам сказал и коэффициент вариации.

Для оценки общей картины колеблемости, для обобщенного представления об амплитудах колебаний нескольких частот коэффициент вариации, может быть, даже предпочтительнее критерия «хи-квадрат», потому что нагляднее показывает различия в силе колебаний и «прямее» их улавливает, к тому же процент вместо отвлеченного числа легче укладывается в нематематическом уме филолога, да и прибегать к услугам таблицы не нужно; чем больше коэффициент, тем больше колеблемость, а критический предел находится где-то около 40%.

Нужно много опытов применения как критерия «хи-квадрат», так и коэффициента вариации, нужны лингвистические и математические обобщения сравнительных достоинств и недостатков того и другого.

## СРАВНЕНИЕ ДОЛЕЙ

Как уже было сказано, статистика дает в руки лингвисту инструменты не только для сравнения частот, но и для сравнения *долей*. Что такое доля? Это отношение наблюдаемой частоты к длине выборки. Формула для вычисления доли аналогична формуле вероятности: $p=\frac{m}{n}$ (в формуле вероятности принято давать прописную, большую букву $P$, в формуле доли — малую, строчную). Иначе говоря, доля — это часть, занимаемая наблюдаемыми фактами в общем их ряду. Например, если была сделана выборка в 1000 знаменательных слов, и в ней оказалось 250 глаголов, то доля глаголов в совокупности всех знаменательных слов равна 0,25, или 25%.

Доли, очевидно, тоже колеблются, как и частоты, около некоторой средней величины, выражая действие закона вероятности. Если колебания долей подчинены одному и тому же (в данных условиях) статистическому закону, они позволяют вычислить *квадратичное отклонение доли*, определяемое формулой $M=\sqrt{\frac{p\cdot q}{n}}$. В этой формуле $p$ — доля изучаемого явления, $q$ — доля всех остальных явлений той же выборки (того же ряда); понятно, что $q$ всегда равно единице минус доля изучаемых явлений, т. е. $q=1-p$. Так, если в нашем примере доля глаголов ($p$) равна 0,25, то доля всех неглаголов ($q$) равна 0,75; $n$ — длина выборки.

Но вернемся к формуле квадратичного отклонения доли. Она подобна формуле квадратичного отклонения частоты — в том смысле, что позволяет уловить некоторые усредненные пределы колебания долей около их средней теоретической величины. Эта формула применяется для сравнения долей одного и того же явления в двух разных статистических совокупностях фактов, например, можно сравнить долю сказуемых среди всех членов предложения в романах Л. Толстого и в его сказках, в прозе Шолохова и Паустовского, в двух главах одного и того же произведения, в художественной прозе и публицистике и т. д.

Для решения таких статистических задач формула квадратичного отклонения получает следующий вид: $\varepsilon_{1,2}=\sqrt{\bar{p}\cdot\bar{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$, здесь $\varepsilon_{1,2}$ — величина квадратичного отклонения средней доли двух сравниваемых совокупностей; $\bar{p}$ и $\bar{q}$ — средние (для двух совокупностей) доли изучаемых явлений и всех остальных; $n_1$ и $n_2$ — размеры выборок.

Предположим, были взяты две текстовые выборки, каждая длиной в 1000 знаменательных слов; в первой выборке

оказалось 200 глаголов, во второй — 150. Можно ли допустить гипотезу о статистическом равенстве долей глаголов в первой и второй выборках, т. е. можно ли допустить, что фактическое различие долей объясняется законами статистического варьирования одной и той же доли (вероятности)?

Применив для решения задачи указанную формулу, мы найдем, что $\varepsilon_{1,2}=0{,}017$. Для получения этого результата мы должны будем предварительно определить $\bar{p}$ и $\bar{q}$. Так как выборки были равны, то $\bar{p}$ мы получим, сложив 0,20 и 0,15 (доли глаголов в двух выборках) и разделив сумму пополам (получится 0,175; следовательно, $\bar{q}=0{,}825$). Полученное числовое значение $\varepsilon_{1,2}$ (т. е. 0,017) нужно сравнить с разностью долей изучаемого явления (в нашем примере — с разностью глаголов) в двух выборках; эта разность в задаче равна 0,05 (0,20 — 0,15 = 0,05). Если квадратичное отклонение доли меньше разности долей втрое или более, мы вправе отвергнуть гипотезу о случайном, т. е. несущественном расхождении долей. То же самое по-иному: если $3\varepsilon \leqslant p_1 - p_2$, то расхождение долей существенно. В нашей задаче квадратичная ошибка средней доли (0,017) меньше разности долей (0,05) почти втрое, даже чуть менее, чем втрое, значит, мы можем отвергнуть гипотезу о случайности расхождения долей (не приняв во внимание очень небольшое превышение разности долей утроенной ошибкой); если бы утроенная квадратичная ошибка заметнее превзошла разность долей, можно было бы сохранить (до повторного, уточняющего опыта) гипотезу о несущественности, случайности того расхождения долей, которое показали выборки.

Ван дер Варден рекомендует для сравнения долей (вероятностей) применять критерий «хи-квадрат» [1]. Одна из формул, пригодных к действию, такова: $\chi^2=\frac{(x_1-n_1\bar{p})^2}{n_1\bar{p}\,\bar{q}}+\frac{(x_2-n_2\bar{p})^2}{n_2\bar{p}\,\bar{q}}$. Здесь $x_1$ и $x_2$ — выборочные частоты, $\bar{p}$ — средняя доля частот, $\bar{q}$ — средняя доля всех остальных элементов в выборках, $n_1$ и $n_2$ — длины, объемы выборок. Если выборки будут равного объема, формулу можно переписать в более простом и знакомом нам виде: $\chi^2=\frac{(x_1-\bar{x})^2}{\bar{x}\cdot\bar{q}}+\frac{(x_2-\bar{x})^2}{\bar{x}\cdot\bar{q}}$.

[1] Б. Л. Ван дер Варден. Математическая статистика. Изд-во иностр. лит. М., 1960, стр. 275—276.



Здесь $x_1$ и $x_2$ — выборочные частоты, $\bar{x}$ — их средняя, получаемая делением их суммы пополам, $\bar{q}$ — выборочная средняя доля всех единиц в выборках, кроме наблюдаемых, т. е. кроме обозначенных $\bar{x}$. Но тут возникает сложный вопрос о числе степеней свободы. Ван дер Варден рекомендует видеть здесь лишь одну степень свободы, т. е. одну свободную, не связанную другими, частоту. Мне же кажется, что в таких случаях нужно видеть две степени свободы, так как в нашей упрощенной формуле свободно может менять свои значения одна из частот ($x_1$ или $x_2$, вторая частота связана средней) и величина $\bar{x}\cdot\bar{q}$, так как $\bar{q}$ не зависит от $x$. Опыты параллельного решения статистических задач — с помощью формулы ошибки средней доли и критерия «хи-квадрат» — говорят о том, что результаты измерения расхождения одних и тех же долей получаются очень неодинаковыми, если формуле «хи-квадрат» мы приписываем одну степень свободы; получается, например, что если в двух равных выборках по 1000 знаменательных слов было насчитано 200 и 160 имен прилагательных, то расхождение долей прилагательных несущественно (говорит формула средней ошибки доли) и то же расхождение существенно (говорит формула «хи-квадрат» при допуске одной степени свободы); если числовое значение «хи-квадрат» применим для двух степеней свободы, показания обеих формул совпадают. Впрочем, это дело математиков, а не лингвистов — определять точные математические условия применения той или иной формулы.

Итак, можно сравнивать частоты, можно сравнивать доли. При сравнении частот мы устанавливаем, случайны или существенны колебания частот около средней, могут или нет наблюдаемые частоты иметь одну и ту же среднюю. При сравнении долей мы решаем аналогичные задачи применительно к вероятности. А в сущности, и сравнение частот, и сравнение долей показывают нам сохранение или нарушение действия статистического закона.

Лингвист в одних случаях может предпочесть сравнение частот, в других — сравнение долей. Например, если предполагается более или менее одинаковое (постоянное) соотношение между одними и другими языковыми единицами или грамматическими категориями в разных текстах и сериях текстов, нужно сравнивать доли (вероятности) и на основе сравнения высказывать статистически досто-

верные гипотезы о сходствах и различиях языковых и речевых стилей, участков языковых структур в разные эпохи и у разных народов и т. д. Так, можно сравнить доли имен и глаголов в стилях, научном и художественном, у Пушкина и Блока, в языках английском и суахили, в русском языке XIV и XX вв. и т. д. Каждое такое сравнение может многое дать при решении самых разнообразных лингвистических исследовательских задач.

Однако, если лингвиста интересует влияние меняющегося содержания текста на выбор им (текстом) из языка единиц одного и того же наименования или интересует реакция языковых единиц и категорий на меняющиеся содержательные условия текста,— нужно сравнивать (вернее, лучше сравнивать, предпочтительнее, удобнее) и оценивать уже не доли, а фактические частоты, наблюдаемые в серии выборок. Оценка методиками и приемами математической статистики таких частот, их колеблемости, позволяет установить случайный или существенный характер наблюдаемых колебаний и сделать на этой основе немало интересных выводов, касающихся взаимовлияния языка и текста; при таком подходе возникает потребность ввести понятия устойчивости и неустойчивости различных элементов языка в речи, устойчивости и неустойчивости речевой структуры текста. Между прочим, знание о том, случайно или существенно отклоняются наблюдаемые частоты от их средней, позволяет лингвисту с большей или меньшей уверенностью выборочные данные переносить на весь текст или совокупность текстов: ведь очевидно, что, чем больше устойчивость частот, чем реже они существенно отклоняются от средней, тем надежнее действие того статистического закона, внешним выявлением которого и оказываются наблюдаемые выборочные частоты с их колебаниями.

## СРАВНЕНИЕ СРЕДНИХ ВЫБОРОЧНЫХ ЧАСТОТ И ЧАСТОТНЫХ РЯДОВ

Помимо сравнения наблюдаемых выборочных частот и долей, лингвист может быть заинтересован и в сравнении средних выборочных частот, например, в тех случаях, когда он хочет изучить языковые единицы и их категории в отключении от меняющихся влияний



содержания текста, когда он хочет перейти с одной, низшей, ступени отвлечения от текста на другую, более высокую, когда нужно перейти от речи к языку, через ряд более или менее сильных усреднений ее, т. е. речи, показателей. Так, очевидно, что разные типы и виды речи, разные языковые и речевые стили удобно характеризовать именно средними частотами и соотношениями таких частот: при этом усредняются частные и местные влияния текста на выбор и применение языковых единиц и остается более или менее постоянная и общая система воздействий, характеризующая тип, стиль и вид речи или видоизменение языка, именуемое его стилем. Правда, все сказанное отнюдь не означает, что для общих статистических характеристик типов и видов речи, стилей языка и т. д. нужны только средние частоты — нет, нужны и оценки колеблемости частот, потому что типы и стили речи характеризуются между прочим и этими колеблемостями, различием и их величины и их статистического характера, их вероятностной специфики, вероятностного качества.

Одним словом, возникает задача сравнения средних частот приемами и средствами математической статистики и задача оценки результатов такого сравнения. Как то и другое делает математическая статистика — разумеется, в элементарных случаях применения ее аппарата?

Итак, конкретная задача: из текстов писателя *A*. было взято 10 выборок по 500 словоупотреблений знаменательных слов; из серии текстов писателя *B*. было сделано столько же выборок такого же объема; интуитивно все выборки писателя *A*. были определены как более или менее однородные; то же самое можно сказать о выборках из текстов писателя *B*. Получены такие числовые данные, характеризующие частоту имен прилагательных: писатель *A*.: 72, 65, 78, 71, 70, 74, 80, 90, 68, 82; писатель *B*.: 80, 93, 84, 83, 78, 67, 85, 86, 75, 89; исследователю нужно узнать, какой характер носит расхождение средних частот — случайно оно или существенно? Математическая статистика дает в руки лингвиста два инструмента для решения задачи. Первый из них называется критерием Стьюдента. Формула этого критерия такова:

$$t=\frac{x_1-\bar{x}_2}{s_{1,2}}\cdot\sqrt{\frac{\kappa_1\cdot\kappa_2}{\kappa_1+\kappa_2}}\,.$$

Здесь $\bar{x}_1$ и $\bar{x}_2$ — сравниваемые средние частоты, $\kappa_1$ и $\kappa_2$ — число выборок (наблюдений) в двух различных сериях; $s_{1,2}$ — несмещенная оценка среднего квадратичного отклонения в двух сериях выборок (об этом показателе разброса частот вокруг средней речь шла в начале книги), вычисляемая для сравнения двух средних частот по формуле:

$$s_{1,2}=\sqrt{\frac{\sum(x_{i1}-\bar{x}_1)^2+\sum(x_{i2}-\bar{x}_2)^2}{\kappa_1+\kappa_2-2}};$$

в этой формуле $x_{i1}$ и $x_{i2}$ — наблюдаемые в первой и второй сериях выборок частоты; остальные символы уже хорошо знакомы [1].

Вернемся к описанной выше задаче. Частоты первого выборочного ряда (из текстов писателя *A*.) дают среднюю частоту, равную 75; частоты второго выборочного ряда (из текстов писателя *B*.) дают среднюю частоту, равную 82. Именно эти величины (75 и 82) мы и должны сопоставить, статистически сравнить при помощи только что показанных формул. Сумма квадратов отклонений выборочных частот от их средней в первом ряду равна 508, во втором — 494. Введем эти величины во вторую из двух показанных формул, т. е. вычислим несмещенную оценку суммы двух средних квадратичных отклонений (в первой и второй сериях выборок); получим $s_{1,2}=\sqrt{\frac{508+494}{10+10-2}}=7{,}5$. Теперь введем эту величину в первую формулу: $t=\frac{82-75}{7{,}5}\times$ $\times\sqrt{\frac{10\cdot10}{10+10}}=0{,}93\cdot2{,}22=2{,}1$. Итак, выборочная величина $t=2{,}1$. Нужно эту величину сравнить с теоретической, табличной. Для этого нам вновь потребуются степени свободы. Их число равно знаменателю под знаком радикала в формуле для вычисления несмещенной оценки суммы квадратичных отклонений, т. е. равно в нашем случае $10+10-2=18$. Поэтому полученную величину сравниваем с величинами в 18-й строке таблицы *t*; находим, что выборочная величина соответствует приблизительно 5%-ной вероятности, эта вероятность не так мала, чтобы гипотезу о равенстве средних отклонить, но и не так велика, чтобы признать ее вполне надежной для сохранения гипотезы о статистическом равенстве двух средних; в таком случае лучше всего осуществить еще один опыт (т. е. взять еще две серии выборок и если новая величина *t* окажется не больше полученной в первый раз, гипотезу можно принять).

[1] О применении формулы Стьюдента см.: Дж. Эдди Юл и М. Дж. Кендэл. Теория статистики. М., 1960, стр. 549—550; А. К. Митропольский. Техника статистических вычислений. М., 1961, стр. 259.



**Извлечение из таблицы числовых значений *t***

| Число степеней свободы | Вероятность большего значения | | | | |
|---|---|---|---|---|---|
| | 0,50 (50%) | 0,20 (20%) | 0,10 (10%) | 0,05 (5%) | 0,025 (2,5%) |
| 1 | 1,000 | 3,078 | 6,314 | 12,706 | 25,452 |
| 2 | 0,816 | 1,886 | 2,920 | 4,303 | 6,205 |
| 3 | 0,765 | 1,638 | 2,353 | 3,182 | 4,176 |
| 4 | 0,741 | 1,533 | 2,132 | 2,776 | 3,495 |
| 5 | 0,727 | 1,476 | 2,015 | 2,571 | 3,163 |
| 6 | 0,718 | 1,440 | 1,943 | 2,447 | 2,969 |
| 7 | 0,711 | 1,415 | 1,895 | 2,365 | 2,841 |
| 8 | 0,706 | 1,397 | 1,860 | 2,306 | 2,752 |
| 9 | 0,703 | 1,383 | 1,833 | 2,262 | 2,685 |
| 10 | 0,700 | 1,372 | 1,812 | 2,228 | 2,634 |
| 11 | 0,697 | 1,363 | 1,796 | 2,201 | 2,593 |
| 12 | 0,695 | 1,355 | 1,782 | 2,179 | 2,560 |
| 13 | 0,694 | 1,350 | 1,771 | 2,160 | 2,533 |
| 14 | 0,692 | 1,345 | 1,761 | 2,145 | 2,510 |
| 15 | 0,691 | 1,341 | 1,753 | 2,131 | 2,490 |
| 16 | 0,690 | 1,337 | 1,746 | 2,120 | 2,473 |
| 17 | 0,689 | 1,333 | 1,740 | 2,110 | 2,458 |
| 18 | 0,688 | 1,330 | 1,734 | 2,101 | 2,445 |
| 19 | 0,688 | 1,328 | 1,729 | 2,093 | 2,433 |
| 20 | 0,688 | 1,325 | 1,725 | 2,086 | 2,423 |
| 21 | 0,687 | 1,323 | 1,721 | 2,080 | 2,414 |
| 22 | 0,686 | 1,321 | 1,717 | 2,074 | 2,406 |
| 23 | 0,685 | 1,319 | 1,714 | 2,069 | 2,398 |
| 24 | 0,685 | 1,318 | 1,711 | 2,064 | 2,391 |
| 25 | 0,684 | 1,316 | 1,708 | 2,060 | 2,385 |

Вероятность большего значения надо, при пользовании этой таблицей, понимать так же, как и вероятность большего значения при пользовании таблицей числовых значений «хи-квадрат». Так, если бы мы в некотором опыте получили, при восемнадцати степенях свободы, *t*, равное

2,5, это означало бы, что гипотезу о статистическом равенстве двух средних надо отвергнуть, потому что в 18-й строке таблицы мы находим величину, очень близкую нашей (2,445), и она соответствует всего 2,5%-ной вероятности большего значения. Иными словами, полученная нами величина необычно редко встречается при равенстве средних, а так как она получена нами с первой попытки, в первом же статистическом испытании текста, то она, вероятнее всего, говорит о том, что две средние частоты, сопоставленные в опыте, отдалены друг от друга на такое расстояние, которое нельзя признать случайным, т. е. возникшим в результате обычного вероятностного колебания,— оно существенно; именно поэтому в первом же испытании текст и выдал нам столь не частое событие, величину $t$, равную 2,5,— не частое не вообще для любых текстов (на языке статистики, для любых совокупностей), а для текстов статистически однородных, подчиненных одному и тому же закону соотношения частот, обладающему общими статистическими свойствами и показателями.

Та же статистическая задача сравнения двух средних частот может решаться по-иному, с помощью так называемого квадратичного отклонения их разности, для вычисления которого рекомендуется формула: $\varepsilon_{1,2}=\sqrt{\frac{\sigma_1^2}{\kappa_1}+\frac{\sigma_2^2}{\kappa_2}}$; в формуле $\sigma_1^2$ и $\sigma_2^2$ — дисперсии двух серий выборок, средние частоты которых сравниваются (вспомним формулу для вычисления дисперсии: $\sigma^2=\frac{\sum(x_i-\bar{x})^2}{\kappa}$; $\kappa_1$ и $\kappa_2$ — количества наблюдений (выборок) в каждой серии. Полученная величина $\varepsilon_{1,2}$ сравнивается с разностью двух средних частот, и если окажется, что эта разность более чем в три раза превосходит ее квадратичное отклонение, гипотеза о несущественности расхождения частот отвергается.

В нашей задаче были средние частоты — 75 и 82; мы уже вычислили суммы квадратов отклонений от средних частот, они были равны соответственно 508 и 494; разделив эти числа на 10 (количество выборок в первой и во второй серии), получим дисперсии: они равны 50,8 и 49,4. Теперь применим нашу новую формулу: $\varepsilon_{1,2}=\sqrt{\frac{50{,}8}{10}+\frac{49{,}4}{10}}=3{,}17$.



Утроив эту величину, получим 9,51, а разность средних равна 7 (82—75), т. е. она менее утроенного квадратичного отклонения; это позволяет сохранить гипотезу о несущественности того расхождения средних, которое дал нам опыт. Таким образом, и первое (с помощью критерия Стьюдента) и второе (с помощью формулы квадратичного отклонения) измерение разности двух средних, предложенных в задаче, дало один и тот же результат: средние отличаются друг от друга несущественно, они разошлись в силу обычного статистического варьирования одной и той же величины, одной и той же вероятности.

Видимо, можно (хотя работы по статистике об этом обычно и не говорят) использовать для сравнения двух выборочных средних частот и «интервалы действительных средних», вычисляемые с помощью формулы ошибки наблюдения (т. е. формулы $L=\frac{t\sigma}{\sqrt{\kappa}}$ или $L=\frac{ts}{\sqrt{\kappa}}$).

Но что такое «интервал действительной средней»? Чтобы освоиться с этим термином, надо задуматься над тем, что полученная в опыте выборочная средняя лишь с известной вероятностью приближается к той «действительной» средней всего изучаемого текста, которую мы не знаем и ради знания которой (приближенного знания) осуществили ряд выборок. О тех ошибках, которые мы могли допустить в оценке величины действительной средней, судя по частотам нескольких выборок, и дает некоторое (тоже приближенное) представление формула $L=\frac{t\sigma}{\sqrt{\kappa}}$. Она показывает те пределы, за которые выход средних частот при повторном изучении текста маловероятен, и, значит, действительная средняя всего текста должна лежать в этих именно пределах.

Вернемся еще раз к задаче, в которой были предложены для сравнения выборочные средние частоты — 75 и 82. Вспомним, что уже вычисленные суммы возведенных в квадрат отклонений от средних были равны 508 и 494. Введя эти величины в формулу среднего квадратичного отклонения $\sigma=\sqrt{\frac{\sum(x_i-\bar{x})^2}{\kappa}}$, получим числа 7,1 и 7,0 (это и есть средние квадратичные отклонения для двух серий выборок). Теперь остается ввести эти числа в формулу ошибки

$L=\frac{t\cdot\sigma}{\sqrt{\kappa}}$, где $t$ примем равным 2,26 (у нас было 10 выборок, значит, использовано 9 степеней свободы, а чтобы обеспечить 95%-ную надежность определения ошибки при девяти степенях свободы, нужно взять $t=2{,}26$).

Вычисляем величину ошибки: для первой серии выборок (в которой средняя равна 75) величина ошибки равна 5,1; для второй серии выборок эта величина равна 5,0. Теперь мы можем определить интервалы действительной средней, обозначив их $\bar{x}_{01}$ и $\bar{x}_{02}$. Первый интервал $\bar{x}_{01}$ получим, прибавив к первой выборочной средней частоте найденную ошибку и вычтя из частоты эту ошибку (ведь ошибиться мы могли и в сторону увеличения, и в сторону уменьшения действительной средней): $75\pm5{,}1=70—80$ (десятой можно пренебречь); значит, интервал действительной средней и первой серии выборок лежит в пределах от 70 до 80; соответственно второй интервал получим, прибавив к 82 пять и отняв от 82 пять (величину ошибки наблюдения); значит, второй интервал лежит в пределах от 77 до 87. Остается сравнить эти интервалы и установить, «накладываются» они друг на друга или нет. Если они «накладываются» (т. е. верхняя граница менее частотного интервала заходит за нижнюю границу более частотного), это говорит о несущественном расхождении средних выборочных частот. В нашей задаче интервалы накладываются один на другой; следовательно, расхождение средних частот было случайным. Таким образом, и третий инструмент сравнения двух средних дал тот же самый ответ.

Применение интервала действительной средней для проверки гипотез о случайном или существенном расхождении двух выборочных средних проще, нежели применение критерия Стьюдента или формулы квадратичной ошибки разности двух средних частот. Однако, по-видимому, интервалы действительной средней дают менее надежные результаты, чем другие два способа проверки гипотез о статистическом равенстве средних частот.

Близким к сравнению средних является сравнение частотных рядов. Вернемся к тем двум частотным рядам, которые вошли в задачу на сравнение средних. Вот эти ряды: а) 72, 65, 78, 70, 74, 80, 90, 68, 82; б) 80, 93, 84, 83, 78, 85, 86, 67, 75, 89. Можно ли каким-то способом установить, принадлежат или нет наши две серии выборок к одной и той же статистической совокупности, т. е. что расхождения частотных рядов случайны и оба ряда рождены, в конце концов, одной и той же вероятностью? Если можно, это значило бы, что, умея сравнивать частотные ряды, мы тем самым, хотя и косвенно, умеем сравнивать и средние частоты (потому что, если существенны или несущественны различия между частотными рядами, это, по-видимому, должно говорить и о существенности или несущественности различий между теми выборочными средними, которые получили выражение в колеблющихся частотных рядах). Математическая статистика имеет специальный и очень неплохо действующий инструмент для сравнения двух частотных рядов, этот инструмент носит название «хи-критерий» и обозначается большой греческой буквой «хи» — X.

Критерий «хи» требует, чтобы частоты двух сравниваемых рядов были объединены в один ранжированный ряд, т. е. такой ряд, в котором частоты расположены в порядке их возрастания (или убывания). В объединенном ранжированном ряду каждая частота ряда А) и каждая частота ряда В) будет занимать свое порядковое место. Каждому месту в особых таблицах соответствует свой числовой показатель со знаком плюс или минус; суммированием показателей, соответствующих порядковым местам одного из сравниваемых частотных рядов, мы получаем некоторую величину, по которой и судим (сравнивая ее с «критической») о существенности или случайности расхождения двух частотных рядов. Но все это лучше показать. Для этого нужна таблица числовых значений $\Psi\left(\frac{r_i}{n+1}\right)$. Далее дана одна из таких таблиц, вычисленная автором книги.

Для пользования этой таблицей нужны дополнительные сведения о критических пределах тех сумм, которые мы будем получать, складывая табличные величины, соответствующие порядковым местам частот одного ряда в общем ранжированном ряду частот (см. таблицу на стр. 48).

Если в опыте сумма числовых значений «пси», соответствующая порядковым местам частот одной серии в общем ранжированном ряду, превзойдет указанное во вспомогательной таблице критическое значение, так называемая «нулевая» гипотеза, т. е. гипотеза о несущественности расхождений между частотными рядами, отклоняется.

Построим ранжированный ряд.

Частоты *А.*: 65, 68, 70, 71, 72, 74, 78, 80, 82, 90.

**Таблица числовых значений $\Psi\left(\frac{r_i}{n+1}\right)$**

| Порядковый номер частоты | $n$ — общее число частот в двух выборочных рядах | | | | |
|---|---|---|---|---|---|
| | $n=10$ | $n=14$ | $n=18$ | $n=20$ | $n=22$ |
| 1 | —1,33 | —1,50 | —1,63 | —1,66 | —1,73 |
| 2 | —0,91 | —1,11 | —1,25 | —1,31 | —1,36 |
| 3 | —0,60 | —0,84 | —1,00 | —1,07 | —1,13 |
| 4 | —0,35 | —0,62 | —0,80 | —0,88 | —0,94 |
| 5 | —0,11 | —0,43 | —0,63 | —0,71 | —0,78 |
| 6 | +0,11 | —0,25 | —0,45 | —0,57 | —0,64 |
| 7 | +0,35 | —0,08 | —0,33 | —0,43 | —0,51 |
| 8 | +0,60 | +0,08 | —0,20 | —0,31 | —0,39 |
| 9 | +0,91 | +0,25 | —0,07 | —0,18 | —0,28 |
| 10 | +1,33 | +0,43 | +0,07 | —0,06 | —0,16 |
| 11 | — | +0,62 | +0,20 | +0,06 | —0,06 |
| 12 | — | +0,84 | +0,33 | +0,18 | +0,06 |
| 13 | — | +1,11 | +0,45 | +0,31 | +0,16 |
| 14 | — | +1,50 | +0,63 | +0,43 | +0,28 |
| 15 | — | — | +0,80 | +0,57 | +0,39 |
| 16 | — | — | +1,00 | +0,71 | +0,51 |
| 17 | — | — | +1,25 | +0,88 | +0,64 |
| 18 | — | — | +1,63 | +1,07 | +0,78 |
| 19 | — | — | — | +1,31 | +0,94 |
| 20 | — | — | — | +1,66 | +1,13 |
| 21 | — | — | — | — | +1,36 |
| 22 | — | — | — | — | +1,73 |

| Разность числа частот в двух выборочных рядах | $n$ — общее число частот в двух выборочных рядах | | | | |
|---|---|---|---|---|---|
| | $n=10$ | $n=14$ | $n=18$ | $n=20$ | $n=22$ |
| 0—1 | 2,60 | 3,11 | 3,63 | 3,86 | 4,08 |
| 2—3 | 2,49 | 3,06 | 3,60 | 3,84 | 4,06 |
| 4—5 | 2,30 | 3,00 | 3,53 | 3,78 | 4,01 |

Частоты *B*.: 67, 75, 78, 80, 83, 84, 85, 86, 89, 93.
Общий ранжированный ряд:

| Место | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Ряд *A* | 65 | — | 68 | 70 | 71 | 72 | 74 | — | 78 | — | — |
| Ряд *B* | — | 67 | — | — | — | — | — | 75 | — | 78 | 80 |
| Место | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | | |
| Ряд *A* | 80 | 82 | — | — | — | — | — | 90 | — | | |
| Ряд *B* | — | — | 83 | 84 | 85 | 86 | 89 | — | 93 | | |

В этой таблице отчетливо видно, какое именно порядковое место занимает в общем ранжированном ряду каждая из частот двух выборочных рядов (затруднения, возникающие при совпадении частот, можно преодолеть путем случайного выбора порядкового места в ранжированном ряду для каждой из двух равных частот).

Теперь нужно из таблицы числовых значений ψ выбрать те, которые соответствуют порядковым местам частот одного выборочного ряда (например, *A*), и определить их сумму. В нашей задаче эта сумма равна 3,79, критическое же значение суммы (смотрим вспомогательную таблицу; общее число выборок было у нас 20 и разность между двумя сериями выборок по их числу нулевая) больше полученного в опыте и равно 3,86. Таким образом, «нулевую» гипотезу, т. е. предположение о несущественности расхождений между двумя частотными рядами, можно принять. Ранее критерий *t*, и формула квадратичного отклонения, и наложение интервалов частот дали нам тот же ответ [1].

Итак, лингвист, пользуясь сравнительно небольшим набором статистических инструментов, может решать большой круг задач на сравнение наблюдаемых частот, средних выборочных частот, частотных рядов и долей. Во всех случаях такого сравнения лингвист ищет ответ на один и тот же, в сущности, вопрос: можно ли наблюдавшееся расхождение частот или долей объяснить действием одной и той же статистической, вероятностной закономерности, ее

[1] О критерии «хи» и его применении см.: Б. Л. Ван дер Варден. Математическая статистика. М., 1960, стр. 346—357 и 419.

случайным варьированием, или же это расхождение надо объяснять действием двух различных вероятностных законов; в первом предположении, если оно подтвердится, будет скрыто убеждение в том, что два текста, давшие две серии выборок, принадлежат к одной и той же «статистической совокупности», они однородны в соотношении изучаемых статистически фактов; во втором предположении, если оно подтвердится, будет заключено уже другое убеждение — в том, что два текста, давшие две серии выборок, принадлежат к двум разным «статистическим совокупностям», они неоднородны по соотношению статистически изучаемых фактов. Этими предположениями и убеждениями будут сразу же поставлены многие вопросы о причинах, приведших к подчинению разных текстов одному вероятностному закону и к их статистической однородности (по изучаемым языковым признакам) или же к подчинению таких текстов разным статистическим, вероятностным законам и к их неоднородности. Но даже и тогда, когда не удастся установить совокупность причин, порождающих или нарушающих статистическую однородность разных текстов, разных типов речи, — и в этих случаях само по себе открытие, описание, обобщение однородности и неоднородности речи будет двигать вперед науку о языке, давая в руки исследователя объективные критерии различения многих еще не установленных закономерностей языкового функционирования и языкового развития; в частности, постепенно будет все яснее вырисовываться объективная — богатейшая и сложная — картина стилевого варьирования языка, его структуры, и стилевого же видоизменения речевых структур; именно в результате широкого обследования приемами статистики самых разных типов и подтипов языка и речи будет со временем получено более глубокое и более точное описание многообразной жизни человеческих языков в их схождениях и расхождениях, в их функционировании и историческом движении.

## ОШИБКИ НАБЛЮДЕНИЯ И ОПРЕДЕЛЕНИЕ ОБЪЕМА ВЫБОРОК ИЗ ТЕКСТА

1. Применение статистических инструментов для изучения языка и речи привлекает внимание лингвистов, в частности, и потому, что позволяет по нескольким выборкам из исследуемого текста (или целой серии текстов), т. е. по его части, судить о нем в целом. Ведь это очень заманчиво — по десяти или двадцати (или даже только пяти) пробам текста в его разных местах построить достаточно аргументированную гипотезу о функционировании языковых явлений в речи романов Л. Толстого, или в речи лирики Н. Некрасова, или в современной газете, или во всей совокупности речевых фактов, обнимаемой термином «язык художественной прозы», «научный стиль», «разговорный тип языка», или в разные периоды развития языка и т. д. Возникает потребность как-то разграничить понятия «статистическое исследование языка» и «статистическое описание». Описание дает регистрацию частот или долей в некотором тексте, скажем, в разных рассказах Чехова и Куприна; такое описание обладает завидной полнотой, оно лишено кажущихся недостатков выборочного изучения, потому что дает наблюдателю не отрывочное, а полное числовое отображение количественных соотношений в тексте. Однако статистическое описание, обладая некоторыми достоинствами, все же не может заменить статистического выборочного исследования, так как не позволяет строить гипотезы, распространяемые исследователем на не изучавшиеся тексты, интуитивно определяемые как однородные изученным. Выборочное же статистическое исследование, не давая, правда, полной картины и этим как будто ограничивая возможности познания законов языка и речи, вместе с тем в действительности намного расширяет возможности лингвиста в таком именно познании. Выборочное статистическое исследование как раз и имеет целью познание законов целого на основе изучения его нескольких частей. Это очень важно, прежде всего, как раз потому, что открывает возможность увидеть закономерности языкового развития и функционирования, а кроме того, это важно и потому, что статистическим описанием можно охватить лишь некоторые сравнительно небольшие тексты, для выборочного же статистического исследования такие границы не поставлены: исследованием можно охватить и прозу Л. Толстого, и поэзию А. Блока, и драматургическую речь А. Н. Островского, и публицистику В. Г. Белинского, и научную речь К. А. Тимирязева, и различные языковые типы или стили в их целостности, и язык разных эпох.

2. Но исследователь языка и речи, решающий применять выборочную статистическую методику, должен каким-

то образом узнавать о тех «ошибках наблюдения», которые неизбежно будут возникать в силу самих вероятностных законов, их специфики, их обязательного варьирования. Лингвист, определивший среднюю выборочную частоту, должен уметь как-то сравнить ее с той «действительной средней», которую он не знает и лишь приближенное значение которой получил на основе выборочного наблюдения. Точно так же лингвист, получивший выборочные доли интересующих его явлений языка, должен уметь каким-то образом сравнить их с «действительной долей», от которой выборочные доли, по-видимому, отклонились.

Математическая статистика дает в руки исследователя особые инструменты, которые позволяют найти так называемую «ошибку наблюдения», т. е. те пределы, в которых может находиться «действительная средняя частота» или «действительная доля», если предположить, что не изучавшиеся участки текста однородны изученным. Об ошибке наблюдения средней частоты однажды уже было сказано, и она была показана в действии.

Вспомним ее несложную формулу: $L=\frac{t\cdot\sigma}{\sqrt{\kappa}}$, здесь $t$ — табличный, теоретический коэффициент, величина которого зависит от числа степеней свободы (т. е. для наблюдателя-лингвиста от количества выборок), $\sigma$ — среднее квадратичное отклонение (или еще лучше пользоваться величиной $s$ — несмещенной оценкой среднего квадратичного отклонения); $\kappa$ — число наблюдений (выборок). Будем думать, что мы уже умеем вычислять среднее квадратичное отклонение (или его несмещенную оценку); на всякий случай напоминаю одну формулу: $s=\sqrt{\frac{\sum(x_i-\bar{x})^2}{\kappa-1}}$. Но $t$ мы вычислять не умеем, нужно обратиться за помощью к специалистам по теории вероятности и математической статистике, к составленным ими таблицам. Вот извлеченные из таких таблиц некоторые данные (см. таблицу на стр. 53).

Теперь мы вооружены для того, чтобы выбрать подходящую величину для коэффициента в формуле ошибки наблюдения. Обычно признается достаточной 95%-ная надежность вычисления средних частот долей и ошибки их наблюдения. Отсюда видно, что если мы имели серию наблюдений из пяти выборок, то нам надо взять $t$, равное 2,78; если выборок было 10, то нужно взять $t$, равное 2,26,



| Количество выборок | Надежность определения ошибки (вероятность) | | | | | |
|---|---|---|---|---|---|---|
| | 99% (0,99) | 97,5 (0,975) | 95% (0,95) | 90% (0,90) | 80% (0,80) | 60% (0,60) |
| 3 | 9,93 | 6,21 | 4,30 | 2,92 | 1,89 | 1,06 |
| 5 | 4,60 | 3,50 | 2,78 | 2,13 | 1,53 | 0,94 |
| 6 | 4,03 | 3,16 | 2,57 | 2,02 | 1,48 | 0,92 |
| 7 | 3,71 | 2,97 | 2,45 | 1,94 | 1,44 | 0,91 |
| 8 | 3,50 | 2,84 | 2,37 | 1,90 | 1,42 | 0,90 |
| 9 | 3,36 | 2,75 | 2,31 | 1,86 | 1,40 | 0,89 |
| 10 | 3,25 | 2,69 | 2,26 | 1,83 | 1,38 | 0,88 |
| 15 | 2,98 | 2,51 | 2,15 | 1,71 | 1,35 | 0,87 |
| 20 | 2,86 | 2,43 | 2,09 | 1,73 | 1,32 | 0,86 |
| 25 | 2,80 | 2,39 | 2,06 | 1,71 | 1,32 | 0,86 |
| 30 | 2,76 | 2,36 | 2,05 | 1,70 | 1,31 | 0,85 |

и т. д. Чем больше коэффициент, тем надежнее результат, т. е. тем вероятнее определяется и ошибка наблюдения и границы действительной средней частоты; с другой стороны, чем больше наблюдений (выборок), тем, в свою очередь, надежнее результаты применения формулы. Однако статистики находят, что в большинстве таких случаев применения формулы ошибки, когда не требуется особо большая точность и надежность (по-видимому, так именно обстоит дело и в статистическом изучении языка и речи), можно брать коэффициент 2, как некоторую постоянную величину, обеспечивающую достаточно надежные результаты при числе выборок десять и более.

Но что значит 95%-ная надежность того или иного коэффициента? Она значит, что вычисленная по формуле ошибка или меньшая встречается (если исходить из выборочных данных о частотах и их колебаниях) примерно 95 раз на сто испытаний текста, подобных тому, которое было осуществлено исследователем всего один раз; отсюда следует, что большая ошибка может встретиться, но всего пять или менее раз на сто испытаний, на сто статистических опытов, аналогичных уже осуществленному. Подобным же образом нужно толковать и другие проценты надежности, хотя такое толкование и не очень, может быть, строго в глазах математика-теоретика. Но для лингвиста, видимо, и оно достаточно.

3. Получив минимальные сведения о формуле ошибки наблюдения, можно перейти к ее использованию в решении лингвистических задач.

Так, вспомним одну из наших задач: даны два ряда частот: а) 72, 65, 78, 71, 70, 74, 80, 90, 68, 82; б) 80, 93, 84, 83, 78, 85, 86, 67, 76, 89. Уже были вычислены суммы квадратов отклонений от средних: для ряда *А* — 508 (при средней частоте 75); для ряда *Б* — 494 (при средней частоте — 82). Мы уже задавали вопрос: какая ошибка в определении средней частоты (вернее, «действительной средней» всего текста, из которого были взяты 10 выборок) была допущена нами, если необследованный текст однороден, по интересующим нас языковым признакам, обследованным его кускам? Применяем формулу ошибки наблюдения; подставив в нее коэффициент, равный 2,26, и среднее квадратичное отклонение, равное 7,1 для ряда *А* и 7 для ряда *Б*, получим $L_1=\frac{2{,}26\cdot 7{,}1}{\sqrt{10}}=5{,}1$; $L_2=\frac{2{,}26\cdot 7}{\sqrt{10}}=5$. Значит, действительная средняя (по данным наших выборок) лежит в ряду *А* в пределах от 69,9 (75 — 5,1) до 80,1 (75 + 5,1), в ряду *Б* в пределах от 77 (82 — 5) до 87 (82 + 5). И можно предполагать 95%-ную надежность наших результатов для текста, однородного тем выборкам, которые изучались. Это значит, что статистика позволяет нам сформулировать гипотезу о том, что и в необследованных кусках текста (или текстов), однородных обследованному, средние частоты не будут выходить из полученных интервалов чаще, чем пять раз на сто опытов; но и эти пять случаев на сто возможны, но не обязательны. Если нам почему-либо потребуется большая надежность, например, в 99%, придется увеличить коэффициент, что повлечет за собой увеличение интервалов «действительных средних частот».

Возьмем еще одну практическую задачу: было сделано по пяти выборок из двух разных текстов, каждая выборка — 500 знаменательных слов. Получены такие частоты имен прилагательных: *А* — 55, 70, 76, 49, 45; *Б* — 52, 78, 88, 22, 25. Каковы ошибки наблюдения и в каких пределах лежат действительные средние частоты? Прежде всего вычислим суммы возведенных в квадрат отклонений каждой фактической частоты от их средней; получим для ряда *А* — 722, для ряда *Б* — 3596, вторая сумма очень велика, и она заметно увеличит ошибку наблюдения. Вычисляем далее средние квадратичные отклонения (или их несмещенные оценки); получаем $\sigma_1=12$, $\sigma_2=27$. Для пяти выборок величину коэффициента $t$, соответствующую 95%-ной надежности, даст нам таблица, — это 2,78. Теперь можно вычислить ошибки: $L_1=\frac{2{,}78\cdot 12}{\sqrt{5}}=14{,}9$; $L_2=\frac{2{,}78\cdot 27}{\sqrt{5}}=33{,}7$. Мы видим, что ошибки, особенно ошибка в определении средней частоты ряда *Б*, значительны. Интервалы действительных средних лежат в пределах: для ряда *А* от 44,1 до 73,9; для ряда *Б* от 19,3 до 86,7. Очевидно, что полученные нами интервалы действительных средних (особенно интервал средней ряда *Б*) очень велики и дают нам слишком неопределенную информацию о действительных средних частотах изучаемых текстов. По-видимому, нужно как-то уменьшить неопределенность информации; это можно сделать или увеличив число выборок, или увеличив размеры каждой выборки (это второе увеличение уменьшит колеблемость, а с нею и среднее квадратичное отклонение). Можно еще уменьшить коэффициент в формуле ошибки наблюдения; но это повлекло бы за собой уменьшение надежности результатов, что также нежелательно.

4. До сих пор мы определяли величину ошибки наблюдения в тех же единицах, которыми измеряется и выборочная средняя частота. Это не всегда удобно, так как не позволяет достаточно наглядно сравнивать величины ошибок: ведь одно дело ошибка в 25 прилагательных при средней частоте в 50 и совсем другое при средней частоте в 500. Вот почему, помимо абсолютной ошибки (мы только что ее получали и применяли), статистика знает еще относительную ошибку. Абсолютная ошибка — это число изучаемых единиц, на которое действительная средняя может быть больше или меньше выборочной средней; относительная ошибка — это отношение абсолютной ошибки к выборочной средней частоте, выраженное в процентах или десятичной дробью. Так, если абсолютная ошибка равна 25, а средняя 50, то относительная ошибка будет равна 0,5, или 50% (25 : 50=0,5); это значит, что абсолютная ошибка составляет одну вторую средней частоты. При той же абсолютной ошибке и средней частоте, равной 500, относительная ошибка становится иной: она равна 0,05, или 5% (25 : 500=0,05), — в этом случае относительная ошибка составляет всего одну двадцатую средней частоты.

Когда вместо абсолютных ошибок мы определяем ошибки относительные, мы получаем возможность точно срав-

нивать их друг с другом, возможность ясно видеть, в каких случаях ошибка велика и в каких мала. В изучении языка и речи методами статистики относительную ошибку в 5—10% можно признать вполне удовлетворительной, а иногда, если условия не позволяют получить такую ошибку, можно пойти и на то, что она окажется равной 15, 20 и даже 30%. Но, конечно, во всех случаях нужно заботиться о том, чтобы ошибка не была слишком большой, т. е. лежала в пределах, близких к 5—10%.

Для вычисления относительных ошибок наблюдения нужно несколько изменить знакомую нам формулу: $\delta = \frac{t\sigma}{\bar{x}\sqrt{\kappa}}$; изменение вида формулы вполне понятно.

5. Если в опыте изучаются не средние частоты, а доли, нужно знать, какую ошибку мы можем допустить в определении «действительной доли» изучаемых фактов во всей их совокупности. Это делается при помощи формул. Вот они: а) абсолютная ошибка доли $L_p = \frac{2\sqrt{pq}}{\sqrt{n}}$, где 2 — постоянный коэффициент, рекомендованный теоретиками-статистиками, $p$ и $q$ — выборочные доли ($p$ — изучаемых фактов, $q$ — всех остальных), $n$ — длина выборки в словах (или других изучаемых единицах языка).

Допустим, мы имеем выборку длиной в 10 000 словоупотреблений знаменательных слов (она может быть составлена из нескольких выборок меньшего объема или может быть не расчлененной на меньшие выборки). В ней оказалось 3500 имен существительных, т. е. их доля равна 0,35. Какова возможная ошибка в определении доли, в каком интервале можно предполагать «действительную долю»? Для применения формулы нужно узнать величину $q$: $1 - 0{,}35 = 0{,}65$. Теперь формула, заполненная конкретными числовыми данными, примет вид: $L_p = 2\sqrt{\frac{0{,}35 \cdot 0{,}65}{10\,000}} = 0{,}0093$. Это значит, что действительная доля может лежать в интервале от $0{,}35 - 0{,}0093$ до $0{,}35 + 0{,}0093$; округлив значение ошибки до сотых, получим интервал действительной доли: 0,34—0,36. Надежность такого ответа приближенно равна 95%.

6. Очевидно, что можно вычислить и относительную ошибку доли. На помощь приходит формула: $\delta_p = \frac{2\sqrt{q}}{\sqrt{pn}}$.



Относительная ошибка доли должна пониматься аналогично относительной ошибке частоты. Относительная ошибка в определении доли — это отношение абсолютной ошибки к величине выборочной доли. В только что решенной задаче доля имен существительных равнялась 0,35; абсолютная ошибка была определена (округленно) как равная 0,01; вычислив отношение 0,01 к 0,35, получим относительную ошибку, она равна 0,029.

Только что мы вычислили вначале абсолютную ошибку, а затем определили ее отношение к выборочной доле, т. е. мы шли кружным путем. Применив формулу относительной ошибки доли, мы прямо получим приблизительно тот же результат (правда, чуть меньше — 0,028 вместо 0,029; это объясняется тем, что в первом, кружном определении относительной ошибки, в вычислениях было допущено округление: вместо 0,0093 было взято 0,01).

Удовлетворительной в решении лингвистических задач можно признать 5—10%-ную относительную ошибку наблюдения; в только что решенной задаче ошибка была очень небольшой — всего около 3%; это, разумеется, лучше, чем 5% и тем более 10%; но можно повторить, что ошибка в 5—10%, т. е. в 0,05—0,10, признается вполне допустимой. Иногда, когда условия опыта не позволяют получить и такую точность, может быть допущена и большая ошибка — в 15—25%. Важно, чтобы ошибка была каждый раз определена и из нее сделаны лингвистические выводы.

7. Формулы абсолютной и относительной ошибки средней частоты и доли позволяют планировать статистический опыт, позволяют определять достаточное число выборок установленного объема или суммарный размер выборки.

Поставим задачу так: нам нужно получить данные о средней частоте глаголов в тексте с вероятностью (надежностью) в 95% и с относительной ошибкой, не превышающей 5%; из предшествующего опыта известно, что среднее квадратичное отклонение глагола в изучаемом тексте приближенно равно 16,5; сколько текстовых выборок нужно взять, если выборочная средняя частота глагола равна 90?

Из формулы относительной ошибки частоты можно получить, путем преобразования, формулу для определения числа наблюдений (выборок): $\kappa = \frac{4\sigma^2}{\delta^2 \cdot \bar{x}^2}$. Введем в эту фор-

мулу данные из условия задачи, т. е. $\sigma=16{,}5$, $\delta=0{,}05$ и $\bar{x}=90$. Получаем: $\kappa=\frac{4\cdot 16{,}5^2}{0{,}05^2\cdot 90^2}=\frac{989}{20{,}25}=49$.

Как видим, нужна большая серия выборок, чтобы получить заданную точность наблюдения. Но предположим, что среднее квадратичное отклонение было не 16,5, а всего 7,5 (это часто бывает в практике изучения языковых явлений). Как изменится ответ нашей формулы? $\kappa=\frac{4\cdot 7{,}5^2}{0{,}05^2\cdot 90^2}=\frac{225}{20{,}25}=11$. Мы видим резкое уменьшение числа выборок!

Но если все же среднее квадратичное отклонение не 7,5, а именно 16,5 и нет возможности сделать серию выборок, измеряемую числом 49? Как быть? Надо пойти на уменьшение точности результатов, например, на то, чтобы допустить относительную ошибку, не в 5, а в 10%. Посмотрим, что даст нам формула при таком изменении задачи: $\kappa=\frac{4\cdot 16{,}5^2}{0{,}10^2\cdot 90^2}=\frac{989}{81}=12$. И в этом случае число выборок заметно падает.

И еще один вариант: допустим, что средняя частота глаголов была не 90, а 110; остальные условия задачи остались, как в первоначальном ее варианте. Что покажет формула? $\kappa=\frac{4\cdot 16{,}5^2}{0{,}05^2\cdot 110^2}=\frac{989}{30{,}25}=32$. Тоже произошло уменьшение числа требуемых выборок, правда, не столь заметное, как при двух предшествующих изменениях в условиях задачи.

Опыт применения статистики для изучения основных явлений морфологии и синтаксиса в разных стилях русского литературного языка XIX—XX вв. убеждает в том, что 10 или 20 выборок длиной в 500 употреблений знаменательных слов каждая, дают вполне удовлетворительную точность наблюдения как средних частот, так и долей; но, конечно, малочастотные явления грамматики и отдельные слова требуют значительно большего числа наблюдений изучаемой частоты или доли.

8. Для планирования числа выборок или их суммарного объема не обязательно применять формулы относительной ошибки. Несколько проще для вычисления формула, построенная на учете абсолютной ошибки. Вот эта формула для определения числа выборок по известной из предшествующего опыта величине среднего квадратичного отклонения и планируемой абсолютной ошибке: $\kappa=\frac{4\sigma^2}{L^2}$.

Из опыта нам известно, что среднее квадратичное отклонение местоимений в изучаемом тексте равно 5,5 при средней выборочной частоте 50 и длине выборки в 500 знаменательных слов. Нужно рассчитать длину серии выборок так, чтобы их было достаточно для получения ошибки наблюдения, не превышающей пяти местоимений. Подумав над задачей, мы поймем, что для формулы не нужны сведения ни о полученной в предшествующем опыте средней частоте, ни о величине выборки. Но косвенно эти сведения полезны, например, для того, чтобы представить величину планируемой ошибки. Вводим в формулу данные из условия задачи: $\kappa=\frac{4\cdot 5{,}5^2}{5^2}=\frac{121}{25}=6$. Оказывается, нужно всего 6 выборок, чтобы получить среднюю с ошибкой, не превышающей 5 единиц. Правда, в нашу формулу молчаливо введен жесткий коэффициент, равный 2 $\left(L=\frac{2\sigma}{\sqrt{\kappa}}\right.$, именно эта формула преобразована в формулу $\left.\kappa=\frac{4\sigma^2}{L^2}\right)$; но мы помним, что надежность определения средней в заданных ошибкой пределах зависит от коэффициента $t$ и что он изменчив; он убывает или увеличивается в зависимости от числа степеней свободы, а значит, в зависимости и от числа наблюдений. Однако, когда мы решаем показанную только что или аналогичную ей задачу, мы как раз не знаем числа выборок; поэтому и приходится принимать $t$, равное 2, так как такая величина коэффициента дает достаточную надежность при 10 и более выборках.

9. Преобразовав формулу определения абсолютной и относительной ошибки доли, можно получить новые формулы, пригодные для определения величины выборки (уже не ряда, серии выборок, а именно объема выборки или выборок, измеряемого количеством слов или иных единиц языка). Вот они: а) формула для определения объема выборки по заданной абсолютной ошибке доли: $n=\frac{4pq}{L^2}$; б) формула для определения объема выборки по заданной относительной ошибке доли: $n=\frac{4q}{\delta^2\cdot p}$.

Применим первую и вторую формулу к решению конкретных задач.

Задача 1-я. Известно из предшествующего опыта, что доля наречий приближенно равна 0,07 (в авторском повествовании и описании в художественной прозе). Какую выборку нужно взять, чтобы абсолютная ошибка доли не превышала 0,005?

$$n = \frac{4 \cdot 0{,}07 \cdot 0{,}93}{0{,}005^2} = 10{,}416.$$

Задача 2-я. Доля наречий та же. Экспериментатор хочет определить ее с относительной ошибкой, не превышающей 0,05. Какой должна быть длина выборки?

$n = \frac{4 \cdot 0{,}93}{0{,}05^2 \cdot 0{,}07} = 21{,}257$ слов (знаменательных, так как доля наречий предварительно устанавливалась в ряду всех слов знаменательных).

Конечно, для практического применения таких расчетов нужно результаты округлять до тысяч или до полутысяч, поэтому ответ на первую задачу — 10 500, на вторую — 21 000.

10. Еще раз вернемся к уже введенным понятиям «надежность определения средней частоты или доли» и «точность такого определения».

Только что предложенные две задачи решены с надежностью в 95%. Это значит, в 95 испытаниях текста из 100 — при условии, что тексты статистически однородны,— запланированная ошибка не будет превзойдена.

Но лучше понятия «надежность» и «точность» еще раз продумать на примерах и формулах определения действительной средней частоты и действительной доли, т. е., иначе говоря, в формулах, позволяющих вычислить ошибку наблюдения. Как мы помним, формула ошибки наблюдения средней частоты имеет в числителе коэффициент, меняющий свое числовое значение — в зависимости от числа степеней свободы и требуемой экспериментатором надежности. Надежность — это вероятность того, что ошибка не превзойдет установленную величину. Если надежность равна 90%, это значит, что мы можем надеяться на то, что указанная формулой ошибка не будет превзойдена в 90 опытах из 100; в десяти же опытах, аналогичных во всем первому, послужившему основанием для вычисления ошибки, она может выйти за установленные пределы. Точность же — это величина ошибки, а еще вернее — величина относительной ошибки. Если надежность говорит нам, как часто при повторении опытов установленная формулой ошибка может превышаться или, наоборот, не будет превышаться, то точность называет нам величину самой ошибки, возможной в таком-то числе случаев из ста (на что указывает уже надежность).



*Есть некоторое закономерное соотношение* между надежностью и точностью: чем больше точность, тем меньше надежность — при тех же размерах ряда выборок или при той же длине суммарной выборки, исчисленной в языковых единицах. Уменьшив точность, мы повышаем надежность, т. е. повышаем нашу уверенность в том, что за указанные пределы средняя частота (или доля) не выйдет при повторных испытаниях текстов, имеющих статистическую структуру, аналогичную изучавшейся в первом опыте. Уменьшая надежность, мы можем получить более точные оценки изучаемых средних частот и долей.

Нельзя, по-видимому, дать никаких жестких рекомендаций об оптимальных соотношениях между надежностью и точностью, к которым должен стремиться исследователь языка и речи. Эти соотношения подсказываются опытом и корректируются результатами применения статистики в языкознании. Можно лишь принять во внимание опыт применения статистики за пределами науки о языке и рекомендации известных статистиков. Этот опыт и эти рекомендации позволяют признать достаточной надежность в 95% и точность в 5—10%. Однако это лишь очень приближенные границы, от которых в процессе статистического исследования можно и нужно отступать в широких пределах, сообразуясь с конкретными условиями эксперимента, структурой текста, уже полученными статистическими данными, возможностью или невозможностью осуществления повторных опытов, соображениями о затратах времени на подсчеты и вычисления и многими иными обстоятельствами.

Обязательной для лингвиста является не та или иная наперед заданная надежность или точность, а соображения научной целесообразности и самый факт установления — на основе данных статистического эксперимента — и надежности и точности в определении средних частот и долей.

11. Лингвисту необходимо свободно ориентироваться в тех данных, которые нужны для планирования величины выборочных серий или суммарного объема выборок (или длины одной выборки).

По-видимому, не очень эффективны не спланированные на основе некоторого предшествующего эксперимента, никак не организованные выборки из текста. Они должны быть хорошо продуманы экспериментатором как в их структурных признаках, так и в их статистических возможностях. В частности, необходимо по возможности строгое, хотя неизбежно интуитивное определение однородности всех выборок в одной и той же серии. Возникает, таким образом, задача предварительной, еще до кодирования и подсчетов, стратификации текста на однородные по языковой структуре речевые пласты или потоки — для того, чтобы все выборки одной и той же серии были взяты из одного и того же потока. Обработка статистических данных либо подтвердит, либо опровергнет интуитивные решения экспериментатора. Обычно эти решения подтверждаются.

Конечно, возможно и такое планирование статистического эксперимента, когда снимается или смягчается влияние меняющегося содержания произведения на речевую структуру, и стратификация текста получает уже иной, обобщенный облик. Этого можно достигнуть либо усреднением единиц подсчета (например, принять за такую единицу не одно знаменательное слово, а пять, или десять), либо увеличением длины каждой выборки, вошедшей в их серию. Чем длиннее выборки, тем меньше сказывается на их языковой структуре влияние меняющегося конкретного содержания произведения, на первый план все отчетливее выступает некая общая и обобщающая статистическая закономерность, которую и улавливает экспериментатор. Правда, при этом утрачивается информация о воздействии конкретного содержания на речевую структуру текста, ослабляется и исчезает возможность узнать, как реагируют те или иные единицы языка на те или иные участки и линии развития конкретного (т. е. логического, эмоционального, психологического, эстетического) содержания текста. И во всех таких случаях остается задача оценки надежности и точности статистических данных и выводов.



## ОРГАНИЗАЦИЯ СТАТИСТИЧЕСКОГО ИЗУЧЕНИЯ ЯЗЫКА И РЕЧИ

1. Уже говорилось о том, что нужно различать статистическое описание языковых элементов и их статистическое исследование. Можно совершенно точно определить частоты всех частей речи и всех грамматических форм различных частей речи в поэтическом рассказе К. Паустовского «Корзина с еловыми шишками». Можно свести полученные в опыте частоты в соответствующие статистические таблицы. Можно словесно описать все полученные данные и высказать некоторые соображения о лингвистическом смысле этих данных. Сами по себе наши данные могут быть интересными для лингвиста, в особенности если их можно будет сопоставить с аналогичными данными, извлеченными из других текстов Паустовского или текстов других авторов.

Однако такие данные не удовлетворяют лингвиста, желающего осуществить статистический эксперимент с целью исследования тех или иных элементов языка в том или ином тексте. Почему? Потому что они недостаточны для обнаружения и понимания тех статистических закономерностей, которым подчинены интересующие нас явления языка в творчестве Паустовского; реальное речевое функционирование элементов языка характеризуется некоторыми колебаниями частот, и сами эти колебания являются одним из объективных признаков речевого стиля; кроме того, исследователь заинтересован в том, чтобы особенности речи рассказа «Корзина с еловыми шишками» поставить в связь с особенностями речи Паустовского как писателя, с особенностями речи других писателей, с особенностями речи художественной и публицистической и т. д. Но для всего этого сплошное описание, осуществляемое на основе подсчета частот в целом тексте, мало пригодно. Требуется применение так называемого выборочного метода.

Что это значит? Лингвист берет из некоторого интересующего его текста (им может быть текст одного произведения, или ряда произведений одного автора, или ряда произведений нескольких авторов) несколько проб, несколько кусков, которые и называет выборками

Естественно, возникает ряд вопросов, связанных с лучшей организацией выборочного изучения языка.

Какого объема, какой длины должны быть выборки? Говоря вообще, они могут быть разного объема — например, от 10 до 10 000 словоупотреблений каждая. Чем меньше выборка, тем легче будут поддаваться частоты интересующих нас явлений влиянию быстро меняющегося содержания текста. Чем активнее интересующие нас факты языка, т. е. чем чаще они применяются, тем меньшие по длине выборки нужны, чтобы проступила изучаемая закономерность. Например, для того чтобы обнаружилась закономерность количественной активности имени существительного в публицистическом тексте, длина одной выборки может быть равна 100 или даже 50 знаменательным словам текста. Но для того чтобы уловить закономерность частотного функционирования отдельного слова (*весна*, *день*, *физик*, *бежать*, *петь*), потребуются выборки в несколько тысяч слов каждая.

Опирающиеся на опыт теоретические соображения позволяют сказать, что для успешного статистического изучения многих явлений морфологии и синтаксиса достаточны и удобны выборки длиной в 500 или даже в 250 знаменательных слов (если изучаются части речи, члены предложения, вообще — не предложения в целом, простые или сложные) или в 250 и даже в 100 самостоятельных предложений (если изучаются предложения в целом).

Нужно ли в выборку включать все слова текста, одно за другим, или же слова брать наугад, по одному, из разных мест произведения?

Если лингвиста интересуют не только сами по себе частоты, но и условия функционирования изучаемых явлений и влияние этих условий на закономерности функционирования языковых элементов,— выборка должна быть сплошной, т. е. должна представлять собой кусок текста установленной длины. Только такие выборки дают надежную информацию и о статистических закономерностях, и о влиянии на них, оказываемом меняющимися условиями текста.

Должны ли быть выборки по возможности однородными или же это не должно беспокоить наблюдателя?

Конечно, выборки должны быть по возможности однородными. Неоднородность текста (жанровая, стилевая, содержательная) даст очень большие колебания частот, их существенные расхождения и тем самым не позволит экспериментатору обнаружить статистическую закономерность. Поэтому нецелесообразно, изучая, предположим, особенности художественной речи Л. Толстого, брать одну выборку из авторского художественного повествования в «Войне и мире», другую — из рассказов для народа, третью — из философских раздумий писателя в романе «Война и мир», а четвертую — из диалога персонажей того же романа. Однородность выборок обычно определяется интуитивно и затем проверяется показаниями статистики.

Какое число выборок может обеспечить достаточно надежные результаты? Об этом уже говорилось. Опыт показывает, что при изучении явлений морфологии и синтаксиса достаточно надежные результаты можно получить, имея 10 выборок длиной в 500 знаменательных слов каждая (или, если изучаются целые предложения,— длиной в 250 предложений). Конечно, увеличение числа выборок до 15 и 20 увеличит и надежность результатов. Иногда же можно ограничиться и пятью — восемью выборками.

Конечно, любая выборка должна быть документирована наблюдателем-лингвистом, т. е. должны быть точно указаны ее границы в тексте. Если какие-то элементы текста (например, прямая речь) не вошли в выборку, это тоже должно быть отмечено наблюдателем.

Можно надеяться, что все более широкое применение статистической методики лингвистами (а затем и литературоведами) будет помогать оптимальному решению многих вопросов организации выборочного исследования языка и речи. Вероятно, подтвердится убеждение, рожденное нашим опытом применения статистики: выборочная методика себя вполне оправдывает, причем предпочтительнее такой ее вариант, при котором все выборки имеют одинаковую длину, — это намного упрощает и ускоряет необходимую статистическую обработку данных, полученных в опыте.

2. Опыт убеждает и в том, что статистическое изучение языка и речи целесообразно вести по определенным программам, каждая из которых представляет собой систематизированные перечни наблюдаемых в тексте языковых единиц или их признаков, причем каждой языковой единице и каждому признаку присваивается свой кодовый номер, или кодовый символ.

Так, если мы намерены изучать соотношение частей речи в разных стилях, мы можем составить такую элементарную программу: имя существительное — 1, имя прилага-

тельное — 2, имя числительное — 3, местоимение — 4, глагол — 5, причастие — 6, деепричастие — 7, наречие — 8, предлог — 9, союз — 0; междометие, модальные слова и частицы остаются такой программой неучтенными. Но программу можно усложнить и ввести в нее все части речи или их внутренние подразделения.

Почему было использовано десять цифр в качестве кодовых знаков? Просто, это удобнее для кодовых записей и их чтения, а кроме того, такой набор цифр — кодовых знаков,— по-видимому, хорош для передачи в будущем статистической первичной обработки кодовых таблиц машинам (не только электронным, но и более простым — счетно-аналитическим).

Имея такую программу и некоторый текст, мы, естественно, можем каждое очередное слово обозначить соответствующим кодовым значком, т. е. одной из цифр, в зависимости от того, к какой части речи встретившееся слово принадлежит. Например, текст «Летят перелетные птицы в осенней дали голубой» получит такое обозначение: 5219212. Такая запись очень экономна и легко читается по тем явлениям и признакам речевой структуры, кодирование которых предусмотрено нашей программой. Но такая запись и ограничивает возможности статистического изучения: ведь за каждой цифрой только один признак, а каждое слово, каждое предложение, даже каждый звук «многопризначен». Для того чтобы расширить возможности программ и кодовых записей, будем обозначать каждую наблюдаемую единицу языка не цифрой, а числом из нескольких цифр, ну, например, пятизначным. Это сразу резко расширит наши возможности фиксирования признаков изучаемых языковых единиц. Можно, скажем, фиксировать принадлежность слова к той или иной части речи (первая цифра числа), его синтаксическую роль (вторая цифра), условия его грамматического господства или грамматической зависимости (третья и четвертая цифры), положение в конце предложений различных типов (пятая цифра). Но что именно и на каком месте записывать, должно быть предусмотрено программой. Поэтому и программа должна получить иной, более сложный вид. Она должна иметь не один ряд признаков, а пять таких рядов, и каждому ряду будет соответствовать порядковое место цифры в числе, обозначающем слово (либо иную единицу языка). Вот один образец усложненной программы.



## Программа № 1. *Части речи русского языка*

| I | | II | | III | |
|---|---|---|---|---|---|
| Имя существительное | — 1 | Подлежащее | — 1 | 1-я позиция зависимости вправо | — 1 |
| Имя прилагательное | — 2 | Сказуемое | — 2 | 2-я позиция зависимости вправо | — 2 |
| Имя числительное | — 3 | Связка | — 3 | 3-я позиция зависимости вправо | — 3 |
| Местоимение | — 4 | Дополнение прямое | — 4 | 4-я позиция зависимости вправо | — 4 |
| Глагол | — 5 | Дополнение косвенное | — 5 | 5-я позиция и далее вправо | — 5 |
| Причастие | — 6 | Определение согласованное | — 6 | 1-я позиция зависимости влево | — 6 |
| Деепричастие | — 7 | Определение несогласованное | — 7 | 2-я позиция зависимости влево | — 7 |
| Наречие | — 8 | Обстоятельство | — 8 | 3-я позиция зависимости влево | — 8 |
| Предлог | — 9 | Вводное слово | — 9 | 4-я позиция зависимости влево | — 9 |
| Союз | — 0 | Обращение | — 0 | 5-я позиция и далее влево | — 0 |

| V | | | |
|---|---|---|---|
| Подчиняет имя существительное | — 1 | В конце простого самостоятельного предложения | — 1 |
| Подчиняет имя прилагательное | — 2 | В конце сложного предложения | — 2 |
| Подчиняет числительное | — 3 | Перед паузой сочинительной | — 3 |
| Подчиняет местоимение | — 4 | Перед паузой подчинительной | — 4 |
| Подчиняет глагол | — 5 | Перед паузой бессоюзной | — 5 |
| Подчиняет причастие | — 6 | Конец придаточного в главном предложении | — 6 |

| IV | | V | |
|---|---|---|---|
| Подчиняет деепричастие | — 7 | Перед обособлением | — 7 |
| Подчиняет наречие | — 8 | В конце обособления | — 8 |
| | | Перед вводной конструкцией | —9 |
| | | В конце вводной конструкции | —0 |

Кодовая запись слов текста в соответствии с этой программой получает такой вид — например, предложение «Летят они в жаркие страны, а я остаюся с тобой» окажется записано так:

| | |
|---|---|
| 52*У*1*У* | 0*УУУУ* |
| 41*У*5*У* | 41*У*5*У* |
| 9*УУУУ* | 52*У*4*У* |
| 267*УУ* | 9*УУУУ* |
| 18123 | 4516*У*2 |

Эта запись требует некоторых пояснений. В программу введено понятие о первой, второй и т. д. позициях зависимости. Условимся считать главные члены предложения грамматически независимыми; тогда любое слово, зависящее непосредственно от подлежащего или сказуемого, мы можем признать стоящим в 1-й позиции зависимости; слово, подчиненное другому слову, имеющему 1-ю позицию зависимости, мы можем признать стоящим во 2-й позиции, и т. д. Можно заметить, что частота слов, находящихся в различных позициях зависимости, своеобразно говорит о структуре предложения и степени ее сложности. Указания «вправо» и «влево» говорят о месте зависимого слова по отношению к грамматически ведущему. Пятый ряд признаков слова имеет в виду его положение в конце схемы предложения — простого или сложного, перед паузами сочинительной связи предложений, подчинительной и бессоюзной их связи, а также перед и после обособленных и вводных оборотов (одиночные вводные слова обозначаются во втором ряду признаков) и одиночных обособлений. Не каждое наблюдаемое слово имеет признаки всех пяти родов. Вот почему появляется дополнительный знак пустоты, отсутствия одного или нескольких признаков; таким знаком может служить любая буква, в только что показанной записи был использован знак *У*.

Ни одна программа не может быть идеальной. Обычно обнаруживается недостаток информации, предусмотренной программой (так кажется лингвисту).



Для усиления программы, для увеличения ее информационной емкости можно использовать дополнительные значки. Так, предложенная выше программа не предусматривает запись сведений о положении сказуемого по отношению к подлежащему. Естественно, это может рассматриваться как недостаток программы — тем более, что третий ряд признаков как раз дает информацию о положении зависимого слова вправо или влево по отношению к ведущему. Как на ходу исправить такой недостаток? Можно ввести в знак сказуемого (цифры 2), через дробную черту, буквы *п* и *л*, и, когда в тексте глагол-сказуемое или имя-сказуемое стоит вправо от подлежащего, писать 52/*п УУУ* или 12/*п УУУ*; когда сказуемое стоит влево от подлежащего, писать соответственно 52/*л УУУ* или 12/*л УУУ*. Можно, конечно, изменить и самое структуру программы, приспособив ее к решению специальной задачи статистического изучения словопорядка в русском языке.

Таким образом, каждому слову соответствует в кодовой записи число из *n* знаков. Порядок цифр в числе соответствует порядку признаковых рядов в программе, этот порядок остается постоянным, пока действует избранная программа. В основу статистического эксперимента кладется, таким образом, отдельное слово, взятое исследователем в тех его признаках, которые определены программой. Никакой иной информации о слове, кроме предусмотренной программой и уточнениями к ней, запись не несет. Но все то, что записано, будет совершенно одинаково расшифровано, прочитано всеми, кто знает программу и код. Таким образом обеспечивается стандартизация статистического эксперимента, независимо от того, кто, где и когда его осуществляет. Это открывает широчайшие возможности осуществления одного и того же эксперимента многими лицами, не связанными общностью места работы и жительства.

Записывать числами слова удобнее так, чтобы запись каждого последующего слова оказывалась строго под записью предшествующего: знаки морфологических признаков должны быть строго друг под другом, также и знаки синтаксических признаков и т. д. Этим обеспечивается легкая обозреваемость однородной информации, записанной кодовыми цифрами: просматривая числа сверху вниз, столбец за столбцом, наблюдатель быстро улавливает наличие в словах тех или иных языковых признаков. Такая запись

столбиками обеспечивает более быстрый подсчет нужных экспериментатору признаков.

Кодирование изучаемых признаков слова (или звука, предложения, словосочетания, морфемы, слога и т. д.) по строго фиксированной программе позволяет передать — со временем — первичную статистическую обработку закодированной информации счетно-аналитическим или электронно-решающим машинам. Это освободит время экспериментатора-лингвиста для выполнения логической работы, необходимой для осмысливания и обобщения статистических данных в плане идей и проблем языкознания.

Однако здесь полезно сказать, что и ручная обработка закодированной языковой информации (вернее, информации об изучаемых свойствах элементов и участков языковой структуры), как и ручное кодирование, дает очень большие результаты при сравнительно небольших затратах времени. Опыт показывает, что студент-дипломник, разрабатывая тему в течение полутора лет (параллельно с выполнением обычных студенческих обязанностей), может записать по пяти-, шестизначной программе 15 000—30 000 знаменательных слов текста, что достаточно для надежных выводов в случаях, когда наблюдаются не очень малые частоты и доли. Кодовая запись по пяти-шестизначной программе 15 000—30 000 слов содержит очень большую информацию о свойствах изучаемых элементов языковой структуры, и обычно студент-дипломник успевает обработать не всю содержащуюся в кодовой записи информацию, а лишь наиболее заметную ее часть. Обычно не удается извлечь из кодовых записей информацию, даваемую сочетаниями признаков,— например, цифрой, указывающей принадлежность слова к определенной части речи, и цифрой, указывающей синтаксическую роль слова. А лингвиста могут и должны интересовать и такие объединения признаков изучаемого слова, как его морфологическая природа и его синтаксическая роль и т. д. Так что для лингвиста, не связанного условиями выполнения студенческой работы, кодовая запись 20 000—30 000 знаменательных слов — это богатейший запас фактов, первично систематизированных и открывающих громадные возможности лингво-статистических и чисто лингвистических обобщений, получаемых под разными углами зрения и в различных аспектах языковой структуры и ее речевого использования. Ведь 20 000 пятизначных кодовых чисел несут информацию о 100 000 признаках 20 000 знаменательных слов. Если же принять во внимание возможность изучения и сочетаний признаков, а также то, что кодируются и служебные, «однопризнаковые» слова, 20 000 пятизначных кодовых записей знаменательных слов содержат информацию не менее чем о 150 000 словесных признаках и их объединений, допускающих статистическое и лингвистическое изучение; соответственно кодовая запись 30 000 знаменательных слов дает информацию не менее чем о 225 000—250 000 признаках кодируемых слов — признаках морфологических, синтаксических и комплексных. Изменение программ, увеличение числа кодируемых каждым числом языковых признаков слова (или предложения) увеличивает, естественно, объем информации, содержащейся в кодовых записях текста.



3. Едва ли целесообразно обсуждать здесь вопрос о типах и лингвистическом содержании различных программ статистического изучения языка и речи. Ответ на этот вопрос должна дать практика применения статистики лингвистами. Но, может быть, было бы полезно упомянуть уже действующие или готовые к действию программы статистического изучения языка и речи, возникшие в опыте применения статистики к решению лингвистических задач, осуществляемом в Горьковском университете.

Первоначально возникла программа, напоминающая недавно показанную в этой главе. Эта программа была предназначена для изучения соотношений частей речи в разных языковых и речевых стилях. Но программа предусматривала и кодирование сведений об основных синтаксических функциях частей речи, и о степени сложности цепей и ветвей членов предложений, и об объеме простых и сложных предложений, и о грамматической природе связей между простыми предложениями в составе сложных.

Эта программа, получившая название программы № 1, уточнялась и видоизменялась, но при всех видоизменениях сохраняла свое главное содержание и назначение. Она позволила получить очень интересные и новые сведения о количественных соотношениях частей речи в русском литературном языке XIX и XX вв. (в стиле художественной литературы), о соотношении сочинительных, подчинительных и бессоюзных связей, о длине простых и сложных предложений, о средней степени сложности структуры простых предложений и т. д. Часть данных, полученных

на основе применения именно этой программы, была опубликована в журнале «Вопросы языкознания» (1965 г., №[illegible]). Программа продолжает действовать, ее возможности только начали осуществляться — применительно к различному текстовому материалу, т. е. разным языковым и речевым стилям разных авторов и разного времени языкового функционирования; желательны дальнейшие опыты применения этой программы для получения новых результатов в изучении еще не опробованных участков речевого потока.

Однако вскоре стало ясно и то, что, помимо обобщенных программ, предусматривающих статистическое изучение языка и речи на некоторых очень абстрактных уровнях и показывающих статистические закономерности, управляющие очень заметными и малоподвижными элементами языковой структуры (части речи, члены предложения, типы предложений и т. д.), нужны иные программы, рассчитанные на изучение языка и речи на менее абстрактных уровнях, при достаточном внимании исследователя к внутреннему членению частей речи, к системам грамматических значений, передаваемых отдельными грамматическими категориями, к многообразию формально-морфематического варьирования категорий, ко всей сложности различных грамматических линий конструирования простых и сложных предложений. Возникли программы изучения структуры простого предложения, структуры сложного предложения, структуры и сочетаемостных свойств имени прилагательного, структуры и сочетаемостных свойств глагола и другие.

Программа № 2 — «Имя существительное» — посвящена обобщению статистической информации о главных грамматических категориях и формах имени существительного, о его функциях в предложении, о его сочетаемости, о его лексико-грамматических признаках.

Эта программа дает обширную и достаточно разнообразную информацию о грамматическом расчленении имени существительного, она позволяет глубже всмотреться в статистические закономерности грамматического строя

Программа № 2. ***Имя существительное***

| I | | II | | III | |
|---|---|---|---|---|---|
| Лексико-грамматические признаки | | Род и число | | Падеж | |
| Конкретное | — 1 | Мужской, единственное | — 1 | Именительный | — 1 |
| Отвлеченное | — 2 | Мужской, множественное | — 2 | Родительный | — 2 |
| Вещественное | — 3 | Средний, единственное | — 3 | Дательный | — 3 |
| Собирательное | — 4 | Средний, множественное | — 4 | Винительный | — 4 |
| Собственное | — 5 | Женский, единственное | — 5 | Творительный | — 5 |
| Прочие | — 6 | Женский, множественное | — 6 | Родительный с предлогом | — 6 |
| | | Прочие | — 7 | Дательный с предлогом | — 7 |
| | | | | Винительный с предлогом | — 8 |
| | | | | Творительный с предлогом | — 9 |
| | | | | Предложный | — 0 |

| IV | | V | | VI | |
|---|---|---|---|---|---|
| Что подчиняет | | Как подчиняет | | Роль в предложении | |
| Имя существительное | — 1 | Вправо, контактно, одно слово | — 1 | Подлежащее | — 1 |
| Имя прилагательное полное | — 2 | Так же, дистантно | — 2 | Сказуемое | — 2 |
| Имя прилагательное краткое | — 3 | Влево, контактно, одно слово | — 3 | Дополнение | — 3 |
| Местоимение | — 4 | Так же, дистантно | — 4 | Определение | — 4 |
| Имя числительное | — 5 | Вправо, контактно, несколько слов | — 5 | Обстоятельство | — 5 |
| Причастие полное | — 6 | Так же, дистантно | — 6 | Дополнение-определение | — 6 |
| Причастие краткое | — 8 | Влево, контактно, несколько слов | — 7 | Дополнение-обстоятельство | — 7 |
| Глагол | — 6 | Так же, дистантно | — 8 | Определение-обстоятельство | — 8 |
| Наречие | — 8 | Влево и вправо, контактно или дистантно | — 9 | Обращение | — 9 |
| Прочие | — 9 | | | Вводный член | — 0 |

русского языка — правда, и на этот раз на достаточно абстрактном уровне, еще не показывающем вариативности различных категорий имени, расчлененности и полисемичности грамматических значений и т. д.

Программа № 3 («Имя прилагательное») посвящена обобщению статистической информации о главных грамматических категориях и формах имени прилагательного, о его функциях в предложении, о некоторых его сочетательных особенностях, о главнейших разновидностях лексического значения прилагательных.

Программа № 4 («Глагол») требует пояснений. Принята следующая классификация залоговых различий глаголов: а) действительный реальный — глагол переходный и при нем винительный падеж объекта; б) действительный потенциальный — глагол прямопереходный, но в тексте винительного падежа со значением объекта не имеет (*в комнате все читали*); в) страдательный реальный — глагол имеет при себе творительный падеж со значением деятеля; г) страдательный потенциальный — глагол несет значение страдательности, но не имеет при себе творительного падежа со значением деятеля (*заводы возводятся во многих городах*; *ботинки с трудом наделись*); д) средний невозвратный — любой глагол, не имеющий признаков действительного или страдательного и не имеющий аффикса *-ся*; е) средний возвратный — любой глагол, не имеющий признаков действительного или страдательного и наделенный аффиксом *-ся*.

## Программа № 3. *Имя прилагательное*

| Лексическое значение | Качественность, относительность | Что подчиняет, чему подчиняется |
|---|---|---|
| Цвета и света — 1 | Качественное, полное — 1 | Имя существительное — 1 |
| Вкуса и запаха — 2 | Качественное, краткое — 2 | Местоимение — 2 |
| Осязания — 3 | Качественное, превосходной степени — 3 | Наречие — 3 |
| | Качественное, сравнительной степени — 4 | Прочие — 4 |
| Звука — 4 | Качественное, | Имени существительному — 5 |
| Размера — 5 | | Местоимению — 6 |
| Отношения к лицу — 6 | | Прочим — 7 |
| Отношения к предмету — 7 | | |



| Лексическое значение | Качественность, относительность | Что подчиняется, чему подчиняется |
|---|---|---|
| Отношения к месту — 8 | в значении относительного — 5 | |
| Отношения ко времени — 9 | Относительное — 6 | |
| Прочие — 0 | Относительное в значении качественного — 7 | |
| | Притяжательное — 8 | |
| | Притяжательное в чужом значении — 9 | |
| | Прочие — 0 | |

IV V VI

| Как подчиняется | Роль в предложении | Обособление |
|---|---|---|
| Вправо, контактно, одно слово — 1 | Подлежащее — 1 | Однословное вправо — 1 |
| Вправо, дистантно, одно слово — 2 | Сказуемое — 2 | Двусловное вправо — 2 |
| Влево, контактно, одно слово — 3 | Определение — 3 | Трехсловное вправо — 3 |
| Влево, дистантно, одно слово — 4 | Дополнение — 4 | Четырехсловное вправо — 4 |
| Вправо, контактно, несколько слов — 5 | Обстоятельство — 5 | Пятисловное вправо — 5 |
| Вправо, ди- | Обращение — 6 | Однословное влево — 6 |
| | Прочие — 7 | Двусловное влево — 7 |
| | | Трехсловное влево — 8 |

| Как подчиняется | Роль в предложении | Обособление |
|---|---|---|
| стантно, несколько слов — 6<br>Влево, контактно, несколько слов — 7<br>Влево, дистантно, несколько слов — 8<br>Вправо и влево несколько слов — 9 | | Четырехсловное влево — 9<br>Пятисловное влево — 0 |

## Программа № 4. *Глагол*

I II III

| Форма | Залог и вид | Время и наклонение |
|---|---|---|
| Инфинитив — 1<br>Причастие полное — 2<br>Причастие краткое — 3<br>Деепричастие — 4<br>1-е лицо — 5<br>2-е лицо — 6<br>3-е лицо — 7<br>Не имеющая лица — 8<br>Прочие — 9 | Действительный реальный, несовершенный — 1<br>Действительный потенциальный, несовершенный — 2<br>Страдательный реальный, несовершенный — 4<br>Страдательный потенциальный, несовершенный — 4<br>Средний, несовершенный — 5<br>Средний, несовершенный, возвратный — 6<br>Действительный | Настоящее — 1<br>Настоящее в значении прошедшего — 2<br>Настоящее в значении будущего — 3<br>Прошедшее — 4<br>Будущее — 5<br>Условное наклонение — 6<br>Повелительное наклонение — 7<br>Изъявительное наклонение в значении условного — 8<br>Изъявительное наклонение в значении повелительного — 9 |



| Форма | Залог и вид | Время и наклонение |
|---|---|---|
| | реальный, совершенный — 7<br>Действительный потенциальный, совершенный — 8<br>Страдательный реальный, совершенный — 9<br>Страдательный потенциальный, совершенный — 0<br>Средний, совершенный — о/н<br>Средний, совершенный, возвратный о/в | Прочие — 0 |

IV V VI

| Что подчиняет | Как подчиняет | Роль в предложении |
|---|---|---|
| Имя существительное — 1<br>Имя прилагательное, полное — 2<br>Имя прилагательное, краткое — 3<br>Местоимение — 4<br>Имя числительное — 5<br>Глагол — 6<br>Причастие — 7<br>Деепричастие — 8<br>Наречие — 9 | Вправо, контактно, одно слово — 1<br>То же, дистантно — 2<br>Влево, контактно, одно слово — 3<br>То же, дистантно — 4<br>Вправо, контактно, несколько слов — 5<br>То же, дистантно — 6<br>Влево, контакт- | Подлежащее — 1<br>Сказуемое — 2<br>Дополнение — 3<br>Определение — 4<br>Обстоятельство — 5<br>Смешан. роль — 6<br>Связка — 7<br>Вводный член — 8 |

| Что подчиняет | Как подчиняет | Роль в предложении |
|---|---|---|
| Прочие — 0 | но, несколько слов — 7<br>То же, дистантно — 8<br>Вправо и влево — 9 | |

4. Программы № 5 и 6 посвящены статистическому изучению синтаксиса — структуры простого и сложного предложений.

## Программа № 5. ***Простое предложение***

| I | II | III |
|---|---|---|
| Модальность | Состав | Части речи |
| Повествовательное, утвердительное — 1<br>Повествовательное, отрицательное — 2<br>Вопросительное, утвердительное — 3<br>Вопросительное, отрицательное — 4<br>Побудительное, утвердительное — 5<br>Побудительное, отрицательное — 6<br>Прочие — 7<br>Связка — 8<br>Междометие — 9<br>Частица — 0 | Двусоставное полное — 1<br>Двусоставное неполное — 2<br>Неопределенно-личное полное — 3<br>Неопределенно-личное неполное — 4<br>Безличное полное — 5<br>Безличное неполное — 6<br>Назывное полное — 7<br>Назывное неполное — 8<br>Инфинитивное полное — 9<br>Инфинитивное неполное — 0 | Имя существительное — 1<br>Имя прилагательное — 2<br>Имя числительное — 3<br>Местоимение — 4<br>Глагол — 5<br>Причастие — 6<br>Деепричастие — 7<br>Наречие — 8<br>Предлог — 9<br>Союз — 0 |



| IV | V | VI |
|---|---|---|
| Члены предложения | Позиции зависимости | Границы членения |
| Подлежащее — 1<br>Сказуемое — 2<br>Прямое дополнение — 3<br>Согласованное определение — 4<br>Другие второстепенные члены — 5<br>Однородные подлежащие — 6<br>Однородные сказуемые — 7<br>Однородные прямые дополнения — 8<br>Однородные согласованные определения — 9<br>Однородные другие второстепенные члены — 0 | Вправо, первая — 1<br>Вправо, вторая — 2<br>Вправо, третья — 3<br>Вправо, четвертая — 4<br>Вправо, пятая и далее — 5<br>Влево, первая — 6<br>Влево, вторая — 7<br>Влево, третья — 8<br>Влево, четвертая — 9<br>Влево, пятая и далее — 0 | Конец любого сложного — 1<br>Конец простого самостоятельного — 2<br>Конец простого перед сочинением — 3<br>То же, перед подчинением — 4<br>Перед бессоюзием — 5<br>Конец придаточного в главном — 6<br>Начало вводного или вставного — 7<br>Конец вводного или вставного — 8<br>Начало обособления — 9<br>Конец обособления — 0 |

Конечно, такая программа легко может быть осложнена введением в нее и других признаков, например: «В составе какого развернутого члена находится слово», «Число слов в предложении или его части» и т. д. Так, очень заманчиво получить сведения о заполнении состава подлежащего, состава сказуемого и состава прямого дополнения; перечень признаков для решения этой задачи мог бы быть таким: в составе подлежащего — 1, в составе сказуемого — 2, в составе прямого дополнения — 3 (для случаев, когда кодируемое слово контактно зависит от подлежащего, сказуемого или прямого дополнения), в составе подлежащего, дистантно — 4, в составе сказуемого, дис-

тантно — 5, в составе прямого дополнения, дистантно — 6. Введение в программу этого ряда признаков потребует, разумеется, уже не шестизначного, а семизначного обозначения каждого слова. Если в программу ввести еще и признаки обособления, запись каждого слова станет восьмизначной. Все это возможно. Но чем больше знаков в кодовой записи каждого отдельного слова, тем сложнее кодирование и расшифровка ручная или машинная. Поэтому на первых порах, может быть, не нужно стремиться к увеличению объема программ за пределы, позволяющие применять пяти-, шестизначные кодовые записи отдельных слов текста.

### Программа № 6. *Сложное предложение*

| I | II | III |
|---|---|---|
| Модальность | Число частей | Тип связи (каждого простого в сложном) |
| Повествовательное, утвердительное — 1 | Одна часть — 1 | Сочинительная — 1 |
| Повествовательное, отрицательное — 2 | Две части — 2 | Бессоюзная — 2 |
| Вопросительное, утвердительное — 3 | Три части — 3 | Главное — 3 |
| Вопросительное, отрицательное — 4 | Четыре части — 4 | Придаточное одиночное — 4 |
| Побудительное, утвердительное — 5 | Пять частей — 5 | Придаточное в неоднородном разветвлении — 5 |
| Побудительное, отрицательное — 6 | Шесть частей — 6 | Придаточное в однородном разветвлении — 6 |
| | Семь частей — 7 | Придаточное во включении — 7 |
| | Восемь частей — 8 | Придаточное во включении и в разветвлении — 8 |
| | Девять частей — 9 | Придаточное в иных типах зависимостей — 9 |
| | Десять частей и более — 0 | |

Хорошо было бы ввести в эту программу и еще один ряд признаков — «Вид придаточного», используя хотя бы традиционное их членение: придаточное подлежащее — 1, придаточное сказуемое — 2, придаточное дополнительное — 3, придаточное определительное — 4, придаточное обстоятельства места — 5, придаточное обстоятельства времени — 6, придаточное обстоятельства причины — 7, придаточное обстоятельства цели — 8, придаточное обстоятельства образа действия — 9, прочие придаточные — 0.

| IV | V | VI |
|---|---|---|
| Соединяющий союз (союзное слово) | Позиция зависимости придаточного | Структура предложения в целом |
| Соединительный — 1 | Вправо, первая — 1 | Конец простого — 1 |
| Противительный — 2 | Вправо, вторая — 2 | Конец сложносочиненного — 2 |
| Разделительный — 3 | Вправо, третья — 3 | Конец сложноподчиненного — 3 |
| Времени — 4 | Вправо, четвертая — 4 | Конец бессоюзного — 4 |
| Места — 5 | Вправо, пятая — 5 | Конец — с сочинением и подчинением — 5 |
| Прочие союзы — 6 | Влево, первая — 6 | Конец — с сочинением и бессоюзной связью — 6 |
| Союзное слово-существительное — 7 | Влево, вторая — 7 | Конец — с подчинением и бессоюзием — 7 |
| Союзное слово-прилагательное — 8 | Влево, третья — 8 | Конец предложения с сочинением, подчинением и бессоюзием — 8 |
| Союзное слово-наречие — 9 | Влево, четвертая — 9 | |
| | Влево, пятая — 0 | |

Выборка для применения этой программы должна включать все предложения, какие есть в тексте; длина же выборки должна, по-видимому, исчисляться только сложными предложениями. Кодовую запись получают все предложения. Признаки программы под номерами I, II, VI относятся ко всему предложению в целом, признаки программы под номерами III, IV, V, VII (если этот ряд войдет в про-

грамму и будет отражать виды придаточных) к отдельным составным частям сложных предложений. Причем нужно иметь в виду некоторые особенности кодирования признаков третьего и четвертого ряда. Признак «сочинительная связь» (или «бессоюзная связь») лучше, удобнее присваивать по следующему, а не предшествующему предложению. Так, если бы пришлось кодировать высказывание «Дождь прошел, ветер утих, и вновь солнце озарило все вокруг», второму предложению мы присвоили бы признак «бессоюзная связь», а третьему — «сочинительная связь» (т. е. соответственно в числах-кодах этих двух предложений, на третьем месте, стояли бы цифры 2 и 1, как то предусмотрено программой). В тех случаях, когда кодируемое предложение одновременно оказывается придаточным и главным, нужно вводить дробь, в соответствии с программой.

Признаки, указанные в четвертом ряду программы, также удобнее приписывать последующему предложению (при сочинении) или предложению придаточному (при подчинении). Признаки второго и шестого ряда в программе соотнесены друг с другом; второй ряд указывает количество частей в сложном, т. е. количество простых, вошедших в состав сложного, а шестой ряд различает — на очень абстрактном уровне — схему предложений; сложение информации, записанной в соответствии с указаниями второго и шестого рядов программы, дает возможность получить обобщенные характеристики количественного состава сложного предложения в зависимости (или независимости) такого состава от общей схемы высказываний.

Пятый ряд программы вводит понятие степени зависимости придаточного от главного. Если придаточное подчинено непосредственно главному, будем считать такое придаточное занимающим первую позицию зависимости; если придаточное *А* подчинено предложению *В*, которое в свою очередь, является придаточным по отношению к главному, будем считать придаточное *А* занимающим вторую позицию зависимости, и т. д. Одним словом, первая, вторая, третья и последующие позиции зависимости определяются (как и в структуре простого предложения) по отношению к грамматически господствующему элементу высказывания.

5. Все предложенные вниманию читателей программы имеют, конечно, опытный, экспериментальный характер, предназначены для широкой и разносторонней проверки, предполагающей их уточнения и изменения.

Но нужно все же сказать, что предварительная проверка этих (и других, не показанных здесь) программ убеждает в том, что и в нынешнем своем виде программы позволяют получать весьма обширную и разнообразную информацию о многих свойствах и признаках грамматической структуры русского языка и русской речи, способную оказать очень заметную помощь лингвистической науке. Думается, что особенно нужна эта помощь в изучении сложных вопросов функционирования грамматической структуры языка, в изучении интересного и глубокого круга проблем, связанных с попытками ученых построить обоснованную теорию стилевой дифференциации языка и речи. Помощь статистики в этой области незаменима ничем. И чем больше будет осуществляться опытов применения статистики в различных ее вариантах (в книге показан лишь один из вариантов статистической методики), тем успешнее будет решать наше языкознание если не все, то очень многие насущные задачи своего движения.

6. Допустим, что лингвист, пользуясь той или иной программой (или без нее, на основе некоторых нефиксированных в программе задач), получил вручную или с участием машины выборочные частоты (именно они являются исходным рубежом всех статистических и лингво-статистических оценок текста). Возникает сразу же задача некоторой экономной и разумной их организации, задача их фиксирования для первоначального и последующих обобщений. Обычно такое фиксирование выборочных частот осуществляется при помощи различных статистических таблиц. Какой можно представить удовлетворительную структуру статистических таблиц, фиксирующих частоты языковых фактов, каким может быть удовлетворительное содержание таких таблиц?

По-видимому, без особой аргументации ясно, что любая статистическая таблица должна хорошо читаться, ее структура и ее содержание должны быть доступны читателю и без особых пояснений ее автора. Это общее требование распространяется, конечно, и на таблицы, показывающие выборочные частоты и их минимально необходимую для лингвиста статистическую обработку.

Что же должно войти в содержание таблицы? Прежде всего, сами выборочные частоты — эта основа основ всех

последующих статистических действий. В таблице полезно иметь также среднюю выборочную частоту, отклонения выборочных частот от их средней, квадраты таких отклонений, их сумму, вычисленные на ее основе «хи-квадрат» и среднее квадратичное отклонение (или его несмещенную оценку), ошибку наблюдения и интервал «действительной средней». Вот схема таблицы такого содержания:

| Выборки | Классы | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Имя существительное | | Имя прилагательное | | Глагол | | Наречие | |
| | $x_i$ | $a_i$ | $x_i$ | $a_i$ | $x_i$ | $a_i$ | $x_i$ | $a_i$ |
| 1-я | 63 | —15 | 13 | —2 | 62 | +10 | 25 | + |
| 2-я | 74 | —4 | 16 | +1 | 53 | +1 | 27 | + |
| 3-я | 72 | —6 | 13 | —2 | 52 | 0 | 25 | + |
| 4-я | 95 | +17 | 19 | +4 | 33 | —14 | 18 | — |
| 5-я | 86 | +8 | 14 | —1 | 55 | +3 | 15 | — |
| $\bar{x}$ | 78 | | 15 | | 52 | | 22 | |
| $\sum a_i^2$ | 630 | | 26 | | 216 | | 108 | |
| $\sigma$ | 7,95 | | 1,61 | | 4,65 | | 3,28 | |
| $L$ | 9,9 | | 2,0 | | 5,5 | | 4,1 | |
| $\chi^2$ | 8,40 | | 1,73 | | 4,15 | | 5,90 | |
| $\bar{x}_0$ | 68—88 | | 13—17 | | 46,5—57,5 | | 18—26 | |

В таблице $x_i$ — выборочные частоты, $\bar{x}$ — средняя выборочная частота, $a_i$ — отклонения выборочных частот от их средней, $\sum a_i^2$ — сумма квадратов отклонений выборочных частот от их средней, $L$ — ошибка наблюдения, вычисленная по формуле $L=\frac{2{,}78\sigma}{\sqrt{\kappa}}$ (в опыте было всего пять выборок, поэтому коэффициент в формуле ошибки наблюдения был взят равным 2,78, чтобы обеспечить надежность около 95%): $\bar{x}_0$ — интервал, в котором можно предполагать «действительную среднюю», вычисляемый по формуле $\bar{x}_0=\bar{x}\pm L$.



Думается, что только что показанная таблица содержит разнообразную и интересную для лингвиста информацию о частотах тех явлений, которые стали объектом статистического эксперимента. Сопоставление таких таблиц, отражающих изучаемые явления языка в текстах разных авторов, разных стилей, разных периодов исторического движения языковых систем, дает возможность сформулировать понимание разнообразных закономерностей функционирования и развития языковой структуры и обосновать применением инструментов статистики сформулированные лингвистические решения и выводы.

Конечно, показанная таблица основана на таком статистическом наблюдении текста, который предполагает равные по длине выборки из однородного статистически речевого потока (однородность, как уже говорилось, определяется интуицией лингвиста). Если выборки не равны по длине, таблица должна получить иную структуру и иное содержание. Но так как лингвист всегда (или почти всегда) может так организовать статистический эксперимент, чтобы выборочные частоты получать из равных по длине выборок, предложенная форма и структура таблицы может, как показывает опыт, найти очень широкое применение.

Особого внимания заслуживают приводимые в статистических таблицах данные, позволяющие на каком-то этапе статистического исследования языка и речи применить инструменты сравнения средних частот, наблюдаемых частот и долей. Такое сравнение совершенно необходимо. Используемое в некоторых работах о языке и речи арифметическое сравнение частот и долей не дает достаточного материала для необходимых выводов, так как арифметическое равенство статистически может оказаться неравенством, а арифметическое неравенство, наоборот, может быть статистическим равенством. В связи с этим таблицы, содержащие арифметические данные (частоты, доли, вычисленные в виде десятичных дробей или процентов), не обработанные статистически, не могут служить основанием для обобщений статистического характера. Приведем для убедительности нашего утверждения один-два примера. В одной из статей, опубликованных во втором выпуске сборника «Вопросы стилистики», издаваемого Саратовским университетом, содержатся в таблице

чисто арифметические данные о частотах сложноподчиненных предложений с одним, двумя и более придаточными; из статьи читатель узнает, что из текстов каждого изучавшегося стиля (художественного, научного, делового и разговорного в двух вариантах) были взяты выборки по 5000 предложений. Арифметические данные не получают никакой статистической обработки, а между тем высказываются гипотезы о больших и меньших частотах; правда, такие гипотезы нередко бывают правильными и обоснованными, но нередко они неправильны, а самое главное, нет никакого доказательства утверждений о равенстве и неравенстве частот.

В статье сказано, что Горький чаще Леонова применял сложноподчиненные предложения в художественном и публицистическом стилях (Горький — 422 и 390, Леонов — 389 и 293). Но ведь частоты 422 и 389 могли случайно отклониться от одной и той же средней,— об этом говорит даже такой строгий критерий, как «хи-квадрат», равный в случае сравнения частот 422 и 390 всего 1,23!

В той же статье утверждается, что у М. Горького больше сложноподчиненных предложений в художественном стиле, чем в публицистике, а в сочинениях К. Федина такие предложения встречаются чаще в публицистике, чем в художественных произведениях:

| Авторы | Стили | |
|---|---|---|
| | художественный | публицистический |
| М. Горький | 422 | 390 |
| К. Федин | 232 | 265 |

Конечно, арифметически 422 больше 390, а 232 меньше 265. Но применим самый строгий критерий проверки нулевой гипотезы, т. е. гипотезы о статистическом равенстве частот 422 и 390, 232 и 265. Мы получим величины «хи-квадрата», равные соответственно 1,23 и 2,32. Если мы вспомним, что «критическая» величина «хи-квадрата», соответствующая 5% вероятности и отклоняющая нулевую гипотезу, равна 3,84, то мы увидим, что статистически нет никаких оснований предполагать, то у М. Горького сложноподчиненные предложения чаще встречаются в художественном стиле и реже в публицистическом, а у К. Федина — чаще в публицистическом и реже в художественном.



Таким образом, как ни заманчивы бывают подчас сами по себе арифметические данные, они недостаточны, чтобы можно было делать обоснованные выводы о соотношении частот одних и тех же языковых явлений в различных условиях их применения. Нужна статистическая проверка предположений лингвиста о равенстве и неравенстве наблюдаемых частот и долей. Поэтому и в статистических таблицах желательно иметь хотя бы самые необходимые оценки арифметически показанных частот и долей — оценки, полученные на основе инструментов математической статистики; если такие оценки не входят в структуру и содержание таблицы, они должны даваться в комментариях к ней. Только при этом условии лингвист имеет право формулировать суждения о равенстве и неравенстве частот и долей, о количественных закономерностях функционирования и развития наблюдаемых в статистическом эксперименте структурных элементов языка.

7. Как уже говорилось, лингвист заинтересован в том, чтобы осуществленный им статистический эксперимент опирался на текстовые выборки равной длины и однородные по составу изучаемых языковых средств. Конечно, однородность до опыта не может быть определена строго. Опыт обнаруживает, насколько соответствует выборочная статистика показаниям интуиции. Поэтому не нужно понимать требование однородности текста слишком жестко. Однако едва ли целесообразно делать выборки одной серии из текстов различных жанров, особенно далеких друг от друга (роман и сказка, лирическое стихотворение и ода и т. д.). Неудачным пришлось бы признать решение лингвиста взять в одну серию выборки из авторской повествовательно-описательной художественной речи и из речи персонажей: та и другая имеют явно различную структуру. Если изучается так называемая разговорная речь по ее отображению в прозе или драме, целесообразно, по-видимому, как-то типизировать персонажей, объединив близких по стилю. В пределах одного большого художественного произведения могут встретиться явно разнородные куски текста: собственно-художественные, публицистические, деловые и даже научные или стилизованные под науку. Конечно, нужно отказаться от выборок, явно от-

клоняющихся по стилю от некоторой художественной нормы изучаемого автора; выборки, от которых пришлось бы отказаться, могли бы составить особую серию, если они оказались бы достаточно однородными.

В выборку могут включаться не все слова текста, а лишь те, которые входят в определенную грамматическую категорию: это могут быть все глаголы, все имена существительные, все имена прилагательные, все наречия, все слова в функции подлежащего, все слова в функции сказуемого, все обособленные члены предложения и т. д. Могут таким же или подобным образом включаться в выборки и предложения определенных грамматических типов.

Такие дифференциальные выборки хороши тогда, когда лингвиста интересуют соотношения частот и долей внутри одного и того же грамматического или лексического типа (части речи, семантической серии слов, разряда звуков, структур одного и того же типа предложения и т. д.). В такую дифференциальную выборку должны войти все факты, охваченные изучаемой грамматической или иной категорией, за исключением тех, которые встретились экспериментатору в кусках текста, нарушающих его однородность. Так, если изучается статистически глагол (в различных его формах, значениях и т. д.), нецелесообразно в выборки — если они берутся из авторской речи — включать глаголы, примененные в речи персонажей.

8. Наблюдаемые в статистическом эксперименте выборочные частоты и выборочные средние частоты (как и доли) должны ясно показываться в статистических таблицах не только ради их статистической оценки, но и ради их лингвистического истолкования.

Если в ряду выборочных частот лингвист замечает резко отклоняющиеся от средней в большую или меньшую стороны, именно такие частоты должны прежде других привлечь внимание. Почему? Потому, что именно они могут нести информацию о тех условиях и причинах, которыми вызывается нарушение некоторой статистической закономерности функционирования изучаемых элементов языковой структуры. Сопоставляя выборки, давшие наибольшие отклонения частот, с остальными, лингвист может обнаружить содержательные причины или, например, стилевые и контекстные условия, вызвавшие нарушение общей закономерности варьирования частот. Таким образом, пристальное внимание к резко отклоняющимся частотам открывает перед лингвистом возможность увидеть и понять такие контекстуальные и ситуативные условия, на которые наиболее заметно реагируют изучаемые структурно-языковые элементы и которые своеобразно характеризуют и изучаемый стиль и особенности функционирования языковых элементов.



Правда, лингвист, пожелавший извлечь интересующую его научную информацию из наблюдения частот, резко отклоняющихся от средней частоты, встретится с трудностями не только лингвистического свойства (необычность задачи и отсутствие опыта ее решения), но и чисто математического (неясность или неизвестность приемов и способов различения «сильно» и «слабо» отклоняющихся частот).

Для преодоления математической трудности можно рекомендовать оценку частотного ряда критерием «хи-квадрат»: если этот критерий сохранит нулевую гипотезу о случайности колебания наблюдаемых частот около их средней, ни одна из частот не была слишком большой; если же критерий «хи-квадрат» нулевую гипотезу отвергнет, наиболее отклонившиеся частоты, по-видимому, и были слишком большими или слишком малыми.

Так, в одном из опытов было взято из авторской речи романов К. Федина десять выборок по 500 знаменательных слов каждая. Были получены следующие частоты глагола: 89, 79, 71, 106, 115, 76, 89, 117, 65, 89; средняя выборочная частота равна 89,6, или округленно 90. Сумма отклонений от средней, возведенных в квадрат, — 2916; $\chi^2 = \frac{2916}{90} = 32{,}4$. Это значение «хи-квадрата» слишком велико, чтобы можно было принять гипотезу о случайности колебания всех частот около средней. Значит, какие-то частоты слишком велики или малы. Естественно предположить, что это те именно частоты, которые дали самые большие отклонения в ту и другую сторону от средней, а именно пятая по порядку частота (она дала отклонение +25), восьмая (+27) и девятая (—25). Оставим в частотном ряду семь членов, отказавшись от наиболее отклонившихся от средней; получим новую среднюю — почти 85,6. Вычислим новую сумму отклонений от средней, возведенных в квадрат, получим около 799,72; деление этой величины на среднюю частоту даст нам величину «хи-квадрат» — около 9,22. Такая величина «хи-квадрат» при шести сте-

пенях свободы вполне разрешает принять гипотезу о слу-
чайности колебания частот около средней. Значит, именно
те три частоты, от которых мы отказались, и были слиш-
ком большими или слишком малыми. И следовательно, лин-
висту небезынтересно посмотреть, чем отличаются по свое-
му содержанию, стилевой окраске, жанровой принадлеж-
кости и т. д. наиболее отклонившиеся от остальной
выборки.

Есть, разумеется, и другие способы узнавания слишко
больших отклонений частот от их средней. Но эти другие
может быть, даже лучшие, чем «хи-квадрат» — спосо
здесь не будут излагаться и обсуждаться.

Возникает задача экспериментального разделения в
борочных частот, например, такого: а) случайные, б) оче
большие, в) очень малые. И очень большие и очень мал
могли бы образовать самостоятельные ряды, каждый
которых мог бы быть оценен при помощи критерия «х
квадрат» или каким-либо иным способом; разумеется,
ряд очень больших и ряд очень малых частот могут ок
заться, в свою очередь, дающими неслучайные колеба
около их средних; это позволило бы выделить еще два,
крайней мере, частотных ряда — сверхбольших и свер
малых частот. Но думается, что практически было бы д
статочным выделение основного ряда частот, ряда больш
частот и ряда малых частот. Если бы это удалось сдела
в изучении того или иного из стилей языка или речи, э
привело бы к построению обоснованной теории структ
но-языковой вероятностной стратификации стилей с во
можностью статистически оценить удельный вес каждо
слоя в стилевом потоке средств языка.

Можно было бы предложить и второй путь стратиф
кации текстов в зависимости от устойчивости или неусто
чивости колебания частот различных языковых элеме
тов. Тексты (а может быть, точнее, речевые структуры
дающие колебания частот определенных элементов язы
в пределах границ существенности величины «хи-квадра
(т. е. от 5 до 95% ее вероятности), вошли бы в основн
стилевой поток или слой; тексты (речевые структуры), да
щие колебания частот в пределах величин критерия «х
квадрат» от границы существенности (5%) до вдвое бол
шей, вошли бы в стилевой поток, или слой, характеризу
мый нарушением статистической устойчивости частот; те
сты (речевые структуры), дающие колебания часто



оцениваемые критерием «хи-квадрат», вдвое или более
превышающим норму, т. е. величину, соответствующую
5%-ной вероятности, вошли бы в стилевой поток, или слой,
характеризуемый большими нарушениями устойчивости.
Так, если для десяти выборок величина «хи-квадрат»,
соответствующая нижней границе существенности (5%-ная
вероятность), равна 16,92, это означало бы, что все речевые
структуры, дающие колебания изучаемых языковых явле-
ний, оцениваемые критерием «хи-квадрат», не превышаю-
щим величину 16,92, были бы отнесены к первому, основ-
ному слою изучаемого стиля или типа речи; если величина
«хи-квадрат» не превзойдет 33,84, речевые структуры,
дающие соответствующие колебания, были бы отнесены
ко второму слою, слою больших колебаний изучаемого яв-
ления; если же оценка колебаний частот даст величину,
превосходящую 33,84, соответствующие речевые структуры
были бы отнесены к третьему статистическому слою изу-
чаемого стиля — слою очень больших колебаний.

Понятно, что первый и второй пути решения задачи ве-
роятностно-статистической стратификации речевых струк-
тур и текстов не вполне совпадают. Первый путь ведет нас
к получению информации о разделении нескольких веро-
ятностных закономерностей, действующих внутри одного
и того же текста; второй же ведет к получению информа-
ции о различении речевых структур и текстов по степени
их устойчивости и неустойчивости.

Думается, что и та и другая информация небезразлич-
на для лингвиста, видящего задачу построения обоснован-
ной теории (или, на первых порах, гипотезы) речевых и
языковых стилей. Правда, статистическая стратификация
речевых структур потребует соответствующих лингвисти-
ческих обоснований, но такие обоснования не могут воз-
никнуть без показаний математической статистики, слу-
жащей лингвистической науке.

## ТАБЛИЦЫ, ОБЛЕГЧАЮЩИЕ СТАТИСТИЧЕСКИЕ ВЫЧИСЛЕНИЯ

1. Не последнее место в ряду причин, вызывающих
осторожное и недоверчивое отношение многих лингвистов
к статистической методике, занимает боязнь большого
объема различных вычислений. Действительно, без вы-

числений, требуемых необходимыми формулами математической статистики, применять статистическую методику нельзя. Чисто арифметические подсчеты частот изучаемых языковых фактов, не проверяемые, не оцениваемые аппаратом математической статистики, мало эффективны. Значит, надо не только считать изучаемые факты — надо еще выполнять некоторые не всегда легкие вычисления, связанные с элементарно-необходимым лингвисту статистическим аппаратом. И тем не менее опасения вычислительной работы необоснованны. Эта работа занимает не больше времени, чем составление картотек, словников, цитат и другие черновые работы, неизбежно осуществляемые при любой методике качественного анализа языка. К тому же далеко не все лингвисты, рискнувшие применить статистику в изучении языка и речи, знают, что существует немало приемов и способов облегчения вычислительной работы, сокращения съедаемого ею времени. Так, умножение и деление можно и нужно передать счетной логарифмической линейке, с ее помощью соответствующие действия над числами осуществляются в несколько раз быстрее, чем при помощи карандаша и бумаги. На счетной линейке можно успешно извлекать квадратные корни и возводить числа во вторую степень — и то и другое действие часто применяется в статистических вычислениях. И одним из главных помощников лингвиста могут стать различные вычислительные таблицы — прежде всего, таблица квадратов чисел и таблица квадратных корней; эти две таблицы можно найти во многих пособиях для высшей и средней школы, и пользоваться этими таблицами умеет всякий, кто окончил десять классов.

2. В статистических вычислениях очень большое место занимает сумма возведенных в квадрат отклонений выборочных частот от их средней. В зависимость от этой суммы поставлены такие величины, как среднее квадратичное отклонение, его несмещенная оценка, критерий «хи-квадрат», ошибка наблюдения и другие. Эту сумму приходится вычислять каждому, кто применяет выборочную статистическую методику. Эту сумму нельзя заранее предусмотреть никакими таблицами — она дается только конкретным статистическим опытом и меняется от одного текста к другому. Поэтому ее приходится вычислять, пользуясь бумагой и карандашом, счетной линейкой, счетами и таблицами квадратов. Что же касается зависимых от этой суммы величин (среднее квадратичное отклонение, несмещенная оценка среднего квадратичного отклонения, ошибка наблюдения и другие), то они могут определяться по особым статистическим таблицам, рассчитать которые можно, применяя показанные ранее формулы и их преобразования.



В этой главе вниманию читателей предлагается таблица, вычисленная автором и предназначенная для облегчения работы, связанной с определением ошибки наблюдения и несмещенной оценки среднего квадратичного отклонения. В таблице даны, с небольшими интервалами, суммы возведенных в квадрат отклонений выборочных частот от их средней и соответствующие этим суммам величины $L$ для пяти, восьми, десяти, пятнадцати, двадцати и двадцати пяти выборок одинаковой длины (объема); дана также величина $s$ для десяти выборок. Пользуясь таблицей, филолог, осуществляющий статистический эксперимент над текстом, вычисляет по данным опыта $\sum a_i^2$ и по таблице находит $L$ и, если нужно, $s$, т. е. ошибку наблюдения и несмещенную оценку среднего квадратичного отклонения.

Таблица № 1.

**Числовые значения $L$ и $S$ в зависимости от величины $\sum a_i^2$**

| $\sum a_i^2$ | $S$ при $K=10$ | $L$ при различном числе выборок | | | | | |
|---|---|---|---|---|---|---|---|
| | | $K=5$ | $K=8$ | $K=10$ | $K=15$ | $K=20$ | $K=25$ |
| 5 | 0,75 | 1,30 | 0,70 | 0,51 | 0,34 | 0,24 | 0,19 |
| 7 | 0,88 | 1,51 | 0,83 | 0,62 | 0,40 | 0,28 | 0,23 |
| 10 | 1,10 | 1,83 | 1,00 | 0,75 | 0,48 | 0,34 | 0,27 |
| 12 | 1,15 | 2,00 | 1,09 | 0,82 | 0,53 | 0,37 | 0,30 |
| 15 | 1,29 | 2,20 | 1,22 | 0,91 | 0,59 | 0,42 | 0,33 |
| 17 | 1,37 | 2,40 | 1,30 | 0,98 | 0,63 | 0,44 | 0,35 |
| 20 | 1,49 | 2,60 | 1,41 | 1,06 | 0,69 | 0,48 | 0,38 |
| 22 | 1,52 | 2,72 | 1,48 | 1,11 | 0,72 | 0,51 | 0,40 |
| 25 | 1,67 | 2,90 | 1,58 | 1,19 | 0,76 | 0,54 | 0,43 |
| 27 | 1,73 | 3,01 | 1,64 | 1,23 | 0,79 | 0,56 | 0,45 |
| 30 | 1,83 | 3,20 | 1,72 | 1,30 | 0,84 | 0,59 | 0,47 |
| 32 | 1,88 | 3,28 | 1,78 | 1,34 | 0,86 | 0,61 | 0,49 |
| 35 | 1,97 | 3,42 | 1,86 | 1,39 | 0,90 | 0,64 | 0,51 |
| 37 | 2,02 | 3,52 | 1,90 | 1,44 | 0,93 | 0,68 | 0,54 |
| 40 | 2,10 | 3,56 | 1,99 | 1,50 | 0,96 | 0,68 | 0,54 |

| $\sum a_i^2$ | $S$ при $K=10$ | $L$ при различном числе выборок | | | | | |
|---|---|---|---|---|---|---|---|
| | | $K=5$ | $K=8$ | $K=10$ | $K=15$ | $K=20$ | $K=25$ |
| 42 | 2,16 | 3,76 | 2,04 | 1,54 | 0,99 | 0,70 | 0,56 |
| 45 | 2,23 | 3,90 | 2,12 | 1,61 | 1,03 | 0,72 | 0,57 |
| 47 | 2,30 | 4,00 | 2,16 | 1,66 | 1,05 | 0,74 | 0,58 |
| 50 | 2,36 | 4,10 | 2,22 | 1,69 | 1,08 | 0,76 | 0,60 |
| 52 | 2,40 | 4,18 | 2,27 | 1,72 | 1,10 | 0,78 | 0,61 |
| 55 | 2,47 | 4,30 | 2,34 | 1,77 | 1,14 | 0,80 | 0,63 |
| 57 | 2,51 | 4,40 | 2,38 | 1,80 | 1,16 | 0,82 | 0,64 |
| 60 | 2,58 | 4,50 | 2,44 | 1,84 | 1,18 | 0,84 | 0,66 |
| 62 | 2,62 | 4,60 | 2,48 | 1,87 | 1,20 | 0,85 | 0,67 |
| 65 | 2,69 | 4,68 | 2,53 | 1,92 | 1,23 | 0,87 | 0,68 |
| 67 | 2,74 | 4,75 | 2,56 | 1,95 | 1,25 | 0,88 | 0,69 |
| 70 | 1,78 | 4,85 | 2,64 | 2,00 | 1,28 | 0,90 | 0,71 |
| 72 | 2,82 | 4,92 | 2,67 | 2,02 | 1,30 | 0,91 | 0,72 |
| 75 | 2,88 | 5,01 | 2,73 | 2,06 | 1,32 | 0,93 | 0,74 |
| 77 | 2,92 | 5,10 | 2,77 | 2,09 | 1,34 | 0,95 | 0,75 |
| 80 | 2,97 | 5,20 | 2,81 | 2,12 | 1,37 | 0,96 | 0,76 |
| 82 | 3,00 | 5,25 | 2,86 | 2,16 | 1,39 | 0,98 | 0,77 |
| 85 | 3,06 | 5,35 | 2,90 | 2,19 | 1,41 | 0,99 | 0,78 |
| 87 | 3,10 | 5,41 | 2,94 | 2,22 | 1,43 | 1,01 | 0,79 |
| 90 | 3,16 | 5,50 | 3,00 | 2,26 | 1,45 | 1,02 | 0,81 |
| 92 | 3,20 | 5,57 | 3,02 | 2,28 | 1,47 | 1,03 | 0,82 |
| 95 | 3,24 | 5,65 | 3,07 | 2,32 | 1,48 | 1,05 | 0,83 |
| 97 | 3,27 | 5,72 | 3,10 | 2,34 | 1,52 | 1,06 | 0,84 |
| 100 | 3,33 | 5,80 | 3,15 | 2,38 | 1,53 | 1,08 | 0,85 |
| 105 | 3,42 | 5,95 | 3,23 | 2,44 | 1,57 | 1,11 | 0,87 |
| 110 | 3,51 | 6,08 | 3,30 | 2,50 | 1,61 | 1,13 | 0,89 |
| 115 | 3,58 | 6,22 | 3,37 | 2,55 | 1,64 | 1,16 | 0,91 |
| 120 | 3,66 | 6,35 | 3,46 | 2,61 | 1,68 | 1,18 | 0,93 |
| 125 | 3,76 | 6,49 | 3,51 | 2,66 | 1,71 | 1,21 | 0,95 |
| 130 | 3,80 | 6,62 | 3,59 | 2,71 | 1,74 | 1,23 | 0,97 |
| 135 | 3,86 | 6,75 | 3,66 | 2,76 | 1,78 | 1,25 | 0,98 |
| 140 | 3,94 | 6,87 | 3,72 | 2,82 | 1,82 | 1,27 | 1,00 |
| 145 | 4,03 | 7,00 | 3,80 | 2,86 | 1,84 | 1,30 | 1,02 |
| 150 | 4,08 | 7,10 | 3,86 | 2,92 | 1,87 | 1,32 | 1,04 |
| 155 | 4,15 | 7,22 | 3,92 | 2,96 | 1,81 | 1,34 | 1,06 |



| $\sum a_i^2$ | $S$ при $K=10$ | $L$ при различном числе выборок | | | | | |
|---|---|---|---|---|---|---|---|
| | | $K=5$ | $K=8$ | $K=10$ | $K=15$ | $K=20$ | $K=25$ |
| 160 | 4,23 | 7,35 | 3,98 | 3,02 | 1,94 | 1,36 | 1,08 |
| 165 | 4,28 | 7,45 | 4,05 | 3,06 | 1,97 | 1,38 | 1,09 |
| 170 | 4,35 | 7,58 | 4,11 | 3,11 | 2,00 | 1,41 | 1,11 |
| 175 | 4,42 | 7,67 | 4,16 | 3,15 | 2,02 | 1,43 | 1,12 |
| 180 | 4,47 | 7,77 | 4,22 | 3,19 | 2,05 | 1,45 | 1,14 |
| 185 | 4,53 | 7,87 | 4,28 | 3,24 | 2,08 | 1,47 | 1,15 |
| 190 | 4,60 | 8,00 | 4,34 | 3,28 | 2,11 | 1,49 | 1,17 |
| 195 | 4,65 | 8,10 | 4,40 | 3,32 | 2,14 | 1,51 | 1,18 |
| 200 | 4,71 | 8,20 | 4,46 | 3,37 | 2,17 | 1,53 | 1,20 |
| 205 | 4,77 | 8,30 | 4,51 | 3,41 | 2,19 | 1,54 | 1,21 |
| 210 | 4,83 | 8,40 | 4,57 | 3,45 | 2,22 | 1,56 | 1,23 |
| 215 | 4,89 | 8,50 | 4,61 | 3,49 | 2,24 | 1,58 | 1,24 |
| 220 | 4,95 | 8,60 | 4,67 | 3,53 | 2,27 | 2,60 | 1,26 |
| 225 | 5,00 | 8,70 | 4,72 | 3,57 | 2,30 | 1,62 | 2,27 |
| 230 | 5,06 | 8,80 | 4,77 | 3,61 | 2,32 | 1,64 | 1,29 |
| 235 | 5,12 | 8,90 | 4,82 | 3,64 | 2,35 | 1,66 | 1,30 |
| 240 | 5,16 | 9,00 | 4,87 | 3,68 | 2,37 | 1,67 | 1,32 |
| 245 | 5,22 | 9,10 | 4,92 | 3,72 | 2,40 | 1,69 | 1,33 |
| 250 | 5,27 | 9,19 | 4,98 | 3,76 | 2,42 | 1,71 | 1,34 |
| 255 | 5,33 | 9,27 | 5,03 | 3,80 | 2,44 | 1,73 | 1,35 |
| 260 | 5,39 | 9,35 | 5,08 | 3,84 | 2,47 | 1,74 | 1,36 |
| 265 | 5,43 | 9,45 | 5,12 | 3,89 | 2,49 | 1,76 | 1,37 |
| 270 | 5,49 | 9,55 | 5,18 | 3,92 | 2,51 | 1,77 | 1,39 |
| 275 | 5,54 | 9,62 | 5,23 | 3,96 | 2,54 | 1,79 | 1,40 |
| 280 | 5,60 | 9,70 | 5,28 | 3,99 | 2,56 | 1,82 | 1,42 |
| 285 | 5,63 | 9,80 | 5,32 | 4,02 | 2,58 | 1,83 | 1,43 |
| 290 | 5,68 | 9,90 | 5,36 | 4,05 | 2,61 | 1,84 | 1,45 |
| 295 | 5,73 | 9,99 | 5,42 | 4,09 | 2,68 | 1,85 | 1,46 |
| 300 | 5,78 | 10,05 | 5,45 | 4,12 | 2,65 | 1,87 | 1,47 |
| 310 | 5,87 | 10,20 | 5,56 | 4,20 | 2,70 | 1,90 | 1,50 |
| 325 | 6,00 | 10,47 | 5,68 | 4,28 | 2,76 | 1,94 | 1,53 |
| 335 | 6,09 | 10,61 | 5,77 | 4,36 | 2,80 | 1,97 | 1,56 |
| 350 | 6,24 | 10,87 | 5,90 | 4,46 | 2,86 | 2,02 | 1,59 |
| 360 | 6,33 | 10,99 | 6,00 | 4,52 | 2,90 | 2,05 | 2,62 |
| 375 | 6,47 | 11,21 | 6,10 | 4,60 | 2,98 | 2,09 | 1,65 |

| $\sum a_i^2$ | $S$ при $K=10$ | $L$ при различном числе выборок | | | | | |
|---|---|---|---|---|---|---|---|
| | | $K=5$ | $K=8$ | $K=10$ | $K=15$ | $K=20$ | $K=25$ |
| 385 | 6,53 | 11,40 | 6,19 | 4,67 | 3,01 | 2,12 | 1,67 |
| 400 | 6,68 | 11,60 | 6,30 | 4,77 | 3,06 | 2,16 | 1,70 |
| 410 | 6,75 | 11,75 | 6,37 | 4,83 | 3,10 | 2,18 | 1,71 |
| 425 | 6,88 | 11,99 | 6,50 | 4,91 | 3,16 | 2,22 | 1,73 |
| 435 | 6,97 | 12,10 | 6,58 | 4,97 | 3,19 | 2,24 | 1,74 |
| 450 | 7,08 | 12,30 | 6,67 | 5,05 | 3,24 | 2,28 | 1,76 |
| 460 | 7,17 | 12,44 | 6,75 | 5,11 | 3,28 | 2,32 | 1,79 |
| 475 | 7,26 | 12,65 | 6,87 | 5,20 | 3,37 | 2,36 | 1,83 |
| 485 | 7,34 | 12,79 | 6,94 | 5,24 | 3,39 | 2,38 | 1,86 |
| 500 | 7,44 | 12,99 | 7,04 | 5,32 | 3,42 | 2,42 | 1,90 |
| 510 | 7,53 | 13,10 | 7,12 | 5,38 | 3,46 | 2,44 | 1,92 |
| 525 | 7,65 | 13,28 | 7,22 | 5,45 | 3,50 | 2,48 | 1,95 |
| 535 | 7,70 | 13,40 | 7,30 | 5,50 | 3,54 | 2,50 | 1,97 |
| 550 | 7,84 | 13,60 | 7,40 | 5,59 | 3,58 | 2,53 | 2,00 |
| 560 | 7,87 | 13,70 | 7,46 | 5,63 | 3,62 | 2,55 | 2,02 |
| 575 | 7,98 | 13,90 | 7,55 | 5,70 | 3,67 | 2,59 | 2,04 |
| 585 | 8,07 | 14,04 | 5,62 | 5,75 | 3,70 | 2,61 | 2,06 |
| 600 | 8,17 | 14,20 | 7,72 | 5,84 | 3,75 | 2,64 | 2,08 |
| 610 | 8,22 | 14,30 | 7,78 | 5,86 | 3,78 | 2,66 | 2,10 |
| 625 | 8,34 | 14,50 | 7,87 | 5,95 | 3,82 | 2,70 | 2,12 |
| 635 | 8,38 | 14,60 | 7,92 | 6,00 | 3,85 | 2,72 | 2,14 |
| 650 | 8,50 | 14,80 | 8,02 | 6,07 | 3,90 | 2,76 | 2,16 |
| 660 | 8,57 | 14,90 | 8,10 | 6,12 | 3,93 | 2,77 | 2,18 |
| 675 | 8,66 | 15,08 | 8,18 | 6,18 | 3,96 | 2,80 | 2,21 |
| 685 | 8,72 | 15,18 | 8,23 | 6,23 | 4,00 | 2,83 | 2,22 |
| 700 | 8,82 | 15,35 | 8,83 | 6,30 | 4,05 | 2,85 | 2,26 |
| 710 | 8,87 | 15,45 | 8,40 | 6,34 | 4,07 | 2,88 | 2,27 |
| 725 | 8,98 | 15,62 | 8,50 | 6,40 | 4,12 | 2,90 | 2,29 |
| 735 | 9,02 | 15,70 | 8,54 | 6,44 | 4,15 | 2,92 | 2,31 |
| 750 | 9,12 | 15,88 | 8,62 | 6,50 | 4,20 | 2,95 | 2,33 |
| 760 | 9,18 | 16,00 | 8,68 | 6,55 | 4,22 | 2,97 | 2,34 |
| 775 | 9,28 | 16,18 | 8,77 | 6,61 | 4,26 | 3,00 | 2,36 |
| 785 | 9,34 | 16,25 | 8,82 | 6,67 | 4,28 | 3,02 | 2,38 |
| 800 | 9,44 | 16,40 | 8,90 | 6,73 | 4,32 | 3,05 | 2,40 |
| 810 | 9,50 | 16,50 | 8,95 | 6,78 | 4,36 | 3,07 | 2,42 |



| $\Sigma a_i^2$ | $S$ при $K=10$ | $L$ при различном числе выборок | | | | | |
|---|---|---|---|---|---|---|---|
| | | $K=5$ | $K=8$ | $K=10$ | $K=15$ | $K=20$ | $K=25$ |
| 825 | 9,58 | 16,65 | 9,05 | 6,85 | 4,40 | 3,10 | 2,44 |
| 835 | 9,64 | 16,75 | 9,10 | 6,89 | 4,43 | 3,12 | 2,46 |
| 850 | 9,72 | 16,90 | 9,20 | 6,92 | 4,46 | 3,15 | 2,48 |
| 860 | 9,76 | 17,00 | 9,24 | 6,98 | 4,48 | 3,17 | 2,49 |
| 875 | 9,86 | 17,15 | 9,31 | 7,04 | 4,52 | 3,20 | 2,51 |
| 885 | 9,90 | 17,25 | 9,35 | 7,08 | 4,55 | 3,22 | 2,53 |
| 900 | 10,00 | 17,40 | 9,43 | 7,13 | 4,58 | 3,24 | 2,55 |
| 910 | 10,04 | 17,50 | 9,50 | 7,17 | 4,62 | 3,25 | 2,56 |
| 925 | 10,12 | 17,65 | 9,59 | 7,20 | 4,65 | 3,28 | 2,58 |
| 935 | 10,18 | 17,75 | 9,65 | 7,27 | 4,68 | 3,30 | 2,60 |
| 950 | 10,24 | 17,90 | 9,70 | 7,33 | 4,71 | 3,33 | 2,62 |
| 960 | 10,30 | 18,00 | 9,76 | 7,37 | 4,74 | 3,34 | 2,63 |
| 975 | 10,38 | 18,15 | 9,84 | 7,43 | 4,78 | 3,36 | 2,65 |
| 985 | 10,45 | 18,23 | 9,90 | 7,48 | 4,80 | 3,38 | 2,67 |
| 1000 | 10,52 | 18,35 | 9,98 | 7,52 | 4,84 | 3,41 | 2,69 |
| 1025 | 10,63 | 18,51 | 10,09 | 7,63 | 4,90 | 3,44 | 2,72 |
| 1050 | 10,80 | 18,80 | 10,20 | 7,70 | 4,95 | 3,49 | 2,75 |
| 1075 | 10,97 | 19,05 | 10,35 | 7,80 | 5,00 | 3,54 | 2,78 |
| 1100 | 11,05 | 19,25 | 10,48 | 7,90 | 5,07 | 3,58 | 2,82 |
| 1125 | 11,19 | 19,45 | 10,60 | 7,98 | 5,14 | 3,62 | 2,85 |
| 1150 | 11,30 | 19,68 | 10,70 | 8,07 | 5,19 | 3,66 | 2,88 |
| 1175 | 11,42 | 19,90 | 10,80 | 8,15 | 5,25 | 3,70 | 2,91 |
| 1200 | 11,57 | 20,02 | 10,90 | 8,23 | 5,30 | 3,74 | 2,94 |
| 1225 | 11,65 | 20,22 | 11,00 | 8,33 | 5,35 | 3,78 | 2,97 |
| 1250 | 11,78 | 20,42 | 11,10 | 8,41 | 5,41 | 3,82 | 3,00 |
| 1275 | 11,88 | 20,62 | 11,22 | 8,50 | 5,47 | 3,86 | 3,03 |
| 1300 | 12,00 | 20,86 | 11,35 | 8,60 | 5,52 | 3,88 | 3,06 |
| 1325 | 12,13 | 21,10 | 11,45 | 8,68 | 5,57 | 3,90 | 3,09 |
| 1350 | 12,25 | 21,28 | 11,58 | 8,75 | 5,62 | 3,96 | 3,12 |
| 1375 | 12,36 | 21,44 | 11,70 | 8,82 | 5,67 | 4,00 | 3,15 |
| 1400 | 12,48 | 21,64 | 11,80 | 8,90 | 5,72 | 4,03 | 3,18 |
| 1425 | 12,56 | 21,90 | 11,90 | 8,98 | 5,77 | 4,08 | 3,21 |
| 1450 | 12,70 | 22,10 | 12,00 | 9,06 | 5,82 | 4,11 | 3,24 |
| 1475 | 12,78 | 22,30 | 12,10 | 9,14 | 5,88 | 4,14 | 3,27 |
| 1500 | 12,90 | 22,45 | 12,20 | 9,20 | 5,92 | 4,18 | 3,29 |

| $\Sigma a_i^2$ | $S$ при $K=10$ | $L$ при различном числе выборок | | | | | |
|---|---|---|---|---|---|---|---|
| | | $K=5$ | $K=8$ | $K=10$ | $K=15$ | $K=20$ | $K=25$ |
| 1525 | 12,98 | 22,62 | 12,30 | 9,30 | 5,95 | 4,22 | 3,31 |
| 1550 | 13,10 | 22,81 | 12,40 | 9,37 | 6,02 | 4,25 | 3,34 |
| 1575 | 13,23 | 23,00 | 12,50 | 9,45 | 6,07 | 4,28 | 3,37 |
| 1600 | 13,30 | 23,20 | 12,60 | 9,52 | 6,12 | 4,32 | 3,40 |
| 1625 | 13,42 | 23,40 | 12,70 | 9,60 | 6,17 | 4,35 | 3,43 |
| 1650 | 13,50 | 23,60 | 12,78 | 9,68 | 6,22 | 4,37 | 3,46 |
| 1675 | 13,62 | 23,75 | 12,88 | 9,75 | 6,27 | 4,42 | 3,49 |
| 1700 | 13,70 | 23,95 | 12,98 | 9,82 | 6,32 | 4,45 | 3,51 |
| 1725 | 13,82 | 24,10 | 13,05 | 9,88 | 6,36 | 4,48 | 3,54 |
| 1750 | 13,90 | 24,30 | 13,15 | 9,96 | 6,40 | 4,52 | 3,56 |
| 1775 | 14,02 | 24,48 | 13,23 | 10,02 | 6,45 | 4,55 | 3,58 |
| 1800 | 14,10 | 24,64 | 13,31 | 10,08 | 6,50 | 4,58 | 3,61 |
| 1825 | 14,23 | 24,82 | 13,40 | 10,14 | 6,55 | 4,62 | 3,64 |
| 1850 | 14,30 | 25,00 | 13,50 | 10,20 | 6,60 | 4,65 | 3,67 |
| 1875 | 14,42 | 25,20 | 13,60 | 10,28 | 6,64 | 4,68 | 3,69 |
| 1900 | 14,50 | 25,40 | 13,70 | 10,37 | 6,68 | 4,70 | 3,71 |
| 1925 | 14,60 | 25,55 | 13,80 | 10,45 | 6,72 | 4,73 | 3,73 |
| 1950 | 14,70 | 25,70 | 13,88 | 10,50 | 6,76 | 4,76 | 3,75 |
| 1975 | 14,80 | 25,85 | 13,98 | 10,56 | 6,80 | 4,80 | 3,77 |
| 2000 | 14,90 | 26,00 | 14,05 | 10,62 | 6,85 | 4,85 | 3,80 |
| 2025 | 15,00 | 26,15 | 14,14 | 10,68 | 6,90 | 4,86 | 3,82 |
| 2050 | 15,10 | 26,30 | 14,23 | 10,75 | 6,93 | 4,89 | 3,85 |
| 2075 | 15,20 | 26,45 | 14,32 | 10,82 | 6,96 | 4,92 | 3,87 |
| 2100 | 15,30 | 26,60 | 14,41 | 10,89 | 7,00 | 4,95 | 3,90 |
| 2125 | 15,38 | 26,75 | 14,50 | 10,95 | 7,04 | 4,98 | 3,92 |
| 2150 | 15,49 | 26,90 | 14,59 | 10,02 | 7,09 | 5,00 | 3,90 |
| 2175 | 15,55 | 27,05 | 14,68 | 11,09 | 7,14 | 5,03 | 3,96 |
| 2200 | 15,65 | 27,20 | 14,77 | 11,16 | 7,19 | 5,07 | 3,99 |
| 2225 | 15,70 | 27,35 | 14,86 | 11,23 | 7,23 | 5,13 | 4,01 |
| 2250 | 15,80 | 27,50 | 14,95 | 11,30 | 7,27 | 5,16 | 4,03 |
| 2275 | 15,88 | 27,65 | 15,03 | 11,36 | 7,31 | 5,16 | 4,05 |
| 2300 | 15,95 | 27,80 | 15,11 | 11,42 | 7,35 | 5,19 | 4,08 |
| 2325 | 16,03 | 27,95 | 15,19 | 11,48 | 7,39 | 5,22 | 4,10 |
| 2350 | 16,12 | 28,10 | 15,27 | 11,54 | 7,43 | 5,25 | 4,13 |
| 2375 | 16,21 | 28,25 | 15,35 | 11,60 | 7,47 | 5,27 | 4,14 |



| $\Sigma a_i^2$ | $S$ при $K=10$ | $L$ при различном числе выборок | | | | | |
|---|---|---|---|---|---|---|---|
| | | $K=5$ | $K=8$ | $K=10$ | $K=15$ | $K=20$ | $K=25$ |
| 2400 | 16,30 | 28,44 | 15,43 | 11,66 | 7,51 | 5,30 | 4,16 |
| 2425 | 16,40 | 28,55 | 15,51 | 11,72 | 7,54 | 5,32 | 4,18 |
| 2450 | 16,50 | 28,70 | 15,59 | 11,78 | 7,57 | 5,34 | 4,20 |
| 2475 | 16,60 | 28,85 | 15,67 | 11,84 | 7,61 | 5,37 | 4,22 |
| 2500 | 16,70 | 29,00 | 15,75 | 11,90 | 7,65 | 5,39 | 4,25 |
| 2525 | 16,77 | 29,14 | 15,82 | 11,96 | 7,69 | 5,42 | 4,27 |
| 2550 | 16,85 | 29,28 | 15,90 | 12,02 | 7,73 | 5,45 | 4,29 |
| 2575 | 16,92 | 29,42 | 15 97 | 12,08 | 7,76 | 5,47 | 4,31 |
| 2600 | 17,00 | 29,56 | 16,05 | 12,14 | 7,80 | 5,50 | 4,33 |
| 2625 | 17,09 | 29,70 | 16,12 | 12,20 | 7,84 | 5,53 | 4,35 |
| 2650 | 17,18 | 29,84 | 16,20 | 12,26 | 7,88 | 5,56 | 4,37 |
| 2675 | 17,25 | 29,98 | 16,28 | 12,32 | 7,92 | 5,58 | 4,39 |
| 2700 | 17,32 | 30,12 | 16,35 | 12,38 | 7,95 | 5,60 | 4,41 |
| 2725 | 17,39 | 30,26 | 16,43 | 12,44 | 7,99 | 5,63 | 4,43 |
| 2750 | 17,46 | 30,40 | 16,50 | 12,50 | 8,02 | 5,65 | 4,45 |
| 2775 | 17,55 | 30,54 | 16,58 | 12,55 | 8,06 | 5,68 | 4,47 |
| 2800 | 17,64 | 30,68 | 16,65 | 12,60 | 8,09 | 5,70 | 4,50 |
| 2825 | 17,72 | 30,82 | 16,73 | 12,65 | 8,13 | 5,73 | 4,52 |
| 2850 | 17,80 | 30,96 | 16,80 | 12,70 | 8,17 | 5,75 | 4,54 |
| 2875 | 17,87 | 31,10 | 16,87 | 12,75 | 8,20 | 5,78 | 4,56 |
| 2900 | 17,94 | 31,24 | 16,95 | 12,80 | 8,24 | 5,80 | 4,58 |
| 2925 | 18,02 | 31,38 | 17,02 | 12,85 | 8,27 | 5,82 | 4,60 |
| 2950 | 18,10 | 31,52 | 17,10 | 12,90 | 8,30 | 5,85 | 4,62 |
| 2975 | 18,16 | 31,66 | 17,17 | 12,95 | 8,33 | 5,88 | 4,63 |
| 3000 | 18,22 | 31,80 | 17,25 | 13,00 | 8,36 | 5,90 | 4,65 |
| 3050 | 18,40 | 32,05 | 17,39 | 13,11 | 8,44 | 5,95 | 4,69 |
| 3100 | 18,52 | 32,30 | 17,53 | 13,22 | 8,52 | 6,00 | 4,73 |
| 3150 | 18,70 | 32,55 | 17,67 | 13,33 | 8,60 | 6,05 | 4,77 |
| 3200 | 18,85 | 32,80 | 17,81 | 13,44 | 8,67 | 6,10 | 4,81 |
| 3250 | 19,00 | 33,05 | 17,95 | 13,55 | 8,74 | 6,15 | 4,85 |
| 3300 | 19,15 | 33,30 | 18,08 | 13,66 | 8,80 | 6,20 | 4,88 |
| 3350 | 19,30 | 33,55 | 18,21 | 13,77 | 8,86 | 6,25 | 4,92 |
| 3400 | 19,42 | 33,80 | 18,34 | 13,88 | 8,91 | 6,29 | 4,96 |
| 3450 | 19,60 | 34,05 | 18,47 | 13,99 | 8,97 | 6,34 | 5,00 |
| 3500 | 19,78 | 34,30 | 18,60 | 14,10 | 9,03 | 6,38 | 5,04 |

| $\Sigma a_i^2$ | $S$ при $K=10$ | $L$ при различном числе выборок | | | | | |
|---|---|---|---|---|---|---|---|
| | | $K=5$ | $K=8$ | $K=10$ | $K=15$ | $K=20$ | $K=25$ |
| 3550 | 19,90 | 34,53 | 18,74 | 14,21 | 9,09 | 6,43 | 5,08 |
| 3600 | 20,02 | 34,76 | 18,87 | 14,31 | 9,16 | 6,47 | 5,12 |
| 3650 | 20,08 | 34,99 | 19,01 | 14,40 | 9,23 | 6,51 | 5,16 |
| 3700 | 20,22 | 35,22 | 19,14 | 14,50 | 9,30 | 5,56 | 5,19 |
| 3750 | 20,38 | 35,45 | 19,28 | 14,60 | 9,37 | 6,60 | 5,22 |
| 3800 | 20,52 | 35,68 | 19,41 | 14,70 | 9,43 | 6,64 | 5,25 |
| 3850 | 20,70 | 35,91 | 19,55 | 14,80 | 9,49 | 6,68 | 5,28 |
| 3900 | 20,30 | 26,14 | 19,68 | 14,90 | 9,54 | 6,72 | 5,31 |
| 3950 | 20,96 | 36,37 | 19,82 | 15,00 | 9,60 | 6,76 | 5,34 |
| 4000 | 21,04 | 36,60 | 19,95 | 15,09 | 9,66 | 6,80 | 5,37 |
| 4050 | 21,20 | 36;83 | 20,07 | 15,18 | 9,72 | 6,84 | 5,41 |
| 4100 | 21,40 | 37,06 | 20,19 | 15,27 | 9,78 | 6,88 | 5,44 |
| 4150 | 21,50 | 37,29 | 20,31 | 15,36 | 9,84 | 6,92 | 5,47 |
| 4200 | 21,63 | 37,52 | 20,43 | 15,45 | 9,90 | 6,97 | 5,51 |
| 4250 | 21,80 | 37,75 | 20,55 | 15,54 | 9,96 | 7,01 | 5,54 |
| 4300 | 21,92 | 37,98 | 20,67 | 15,63 | 10,02 | 7,05 | 5,57 |
| 4350 | 22,02 | 38,21 | 20,79 | 15,72 | 10,08 | 7,10 | 5,60 |
| 4400 | 22,18 | 38,44 | 20,91 | 15,81 | 10,14 | 7,14 | 5,64 |
| 4450 | 22,28 | 38,67 | 21,03 | 15,90 | 10,19 | 7,20 | 5,67 |
| 4500 | 22,40 | 38,90 | 21,15 | 15,99 | 10,25 | 7,25 | 5,70 |
| 4550 | 22,52 | 39,11 | 21,27 | 16,08 | 10,31 | 7,29 | 5,74 |
| 4600 | 22,64 | 39,32 | 21,39 | 16,17 | 10,37 | 7,33 | 5,76 |
| 4650 | 22,78 | 39,53 | 21,50 | 16,26 | 10,42 | 7,37 | 5,82 |
| 4700 | 22,94 | 39,74 | 21,62 | 16,35 | 10,47 | 7,40 | 5,85 |
| 4750 | 23,04 | 39,95 | 21,73 | 16,43 | 10,53 | 7,44 | 5,89 |
| 4800 | 23,18 | 40,16 | 21,84 | 16,51 | 10,59 | 7,47 | 5,92 |
| 4850 | 23,24 | 40,37 | 21,96 | 16,60 | 10,64 | 7,51 | 5,95 |
| 4900 | 23,38 | 40,58 | 22,07 | 16,68 | 10,70 | 7,55 | 5,99 |
| 4950 | 23,46 | 40,79 | 22,19 | 16,76 | 10,75 | 7,59 | 6,02 |
| 5000 | 23,62 | 41,00 | 22,30 | 16,85 | 10,80 | 7,63 | 6,05 |
| 5050 | 23,70 | 41,20 | 22,41 | 16,93 | 10,86 | 7,67 | 6,08 |
| 5100 | 23,82 | 41,40 | 22,52 | 17,01 | 10,91 | 7,70 | 6,11 |
| 5150 | 23,95 | 41,60 | 22,63 | 17,09 | 10,97 | 7,74 | 6,14 |
| 5200 | 24,02 | 41,80 | 22,74 | 17,17 | 11,02 | 7,78 | 6,16 |
| 5250 | 24,08 | 42,00 | 22,85 | 17,25 | 11,09 | 7,81 | 6,19 |



| $\Sigma a_i^2$ | $S$ при $K=10$ | $L$ при различном числе выборок | | | | | |
|---|---|---|---|---|---|---|---|
| | | $K=5$ | $K=8$ | $K=10$ | $K=15$ | $K=20$ | $K=25$ |
| 5300 | 24,22 | 42,20 | 22,96 | 17,33 | 11,14 | 7,85 | 6,21 |
| 5350 | 24,40 | 42,40 | 23,07 | 17,41 | 11,20 | 7,89 | 6,23 |
| 5400 | 24,54 | 42,60 | 23,18 | 17,49 | 11,25 | 7,93 | 6,26 |
| 5450 | 24,62 | 42,80 | 23,29 | 17,57 | 11,30 | 7,97 | 6,28 |
| 5500 | 24,72 | 43,00 | 23,40 | 17,65 | 11,35 | 8,00 | 6,30 |
| 5550 | 24,82 | 43,19 | 23,50 | 17,73 | 11,40 | 8,04 | 6,33 |
| 5600 | 24,92 | 43,38 | 23,60 | 17,81 | 11,45 | 8,07 | 6,36 |
| 5650 | 25,02 | 43,57 | 23,70 | 17,89 | 11,50 | 8,11 | 6,39 |
| 5700 | 25,10 | 43,76 | 23,80 | 17,97 | 11,55 | 8,15 | 6,42 |
| 5750 | 25,22 | 43,95 | 23,90 | 18,05 | 11,60 | 8,19 | 6,45 |
| 5800 | 25,36 | 44,14 | 24,00 | 18,13 | 11,65 | 8,22 | 6,48 |
| 5850 | 25,50 | 44,33 | 24,10 | 18,21 | 11,70 | 8,26 | 6,51 |
| 5900 | 25,60 | 44,52 | 24,20 | 18,29 | 11,75 | 8,29 | 6,54 |
| 5950 | 25,74 | 44,71 | 24,30 | 18,37 | 11,80 | 8,33 | 6,57 |
| 6000 | 25,84 | 44,90 | 24,40 | 18,45 | 11,85 | 8,36 | 6,60 |
| 6050 | 25,98 | 45,08 | 24,50 | 18,53 | 11,90 | 8,39 | 6,63 |
| 6100 | 26,08 | 45,26 | 24,60 | 18,61 | 11,95 | 8,43 | 6,65 |
| 6150 | 26,18 | 45,44 | 24,70 | 18,69 | 12,00 | 8,46 | 6,67 |
| 6200 | 26,28 | 45,62 | 24,80 | 18,77 | 12,05 | 8,49 | 6,70 |
| 6250 | 26,38 | 45,80 | 24,90 | 18,85 | 12,10 | 8,53 | 6,73 |
| 6300 | 26,44 | 45,98 | 25,00 | 18,92 | 12,15 | 8,56 | 6,75 |
| 6350 | 26,54 | 46,16 | 25,10 | 18,99 | 12,20 | 8,59 | 6,78 |
| 6400 | 26,64 | 46,34 | 25,20 | 19,06 | 12,25 | 8,63 | 6,80 |
| 6450 | 26,74 | 46,52 | 25,30 | 19,13 | 12,30 | 8,67 | 6,83 |
| 6500 | 26,84 | 46,70 | 25,40 | 19,20 | 12,35 | 8,70 | 6,85 |
| 6550 | 26,98 | 46,88 | 25,50 | 19,27 | 12,40 | 8,73 | 6,88 |
| 6600 | 27,08 | 47,06 | 25,60 | 19,34 | 12,45 | 8,77 | 6,91 |
| 6650 | 27,18 | 47,24 | 25,70 | 19,41 | 12,50 | 8,80 | 6,93 |
| 6700 | 27,28 | 47,42 | 25,80 | 19,48 | 12,55 | 8,83 | 6,95 |
| 6750 | 27,38 | 47,60 | 25,90 | 19,55 | 12,59 | 8,87 | 6,97 |
| 6800 | 27,48 | 47,78 | 26,00 | 19,62 | 12,64 | 8,90 | 7,00 |
| 6850 | 27,58 | 47,96 | 26,10 | 19,69 | 12,68 | 8,93 | 7,02 |
| 6900 | 27,68 | 48,14 | 26,20 | 19,76 | 12,72 | 8,96 | 7,05 |
| 6950 | 27,78 | 48,32 | 26,30 | 19,83 | 12,76 | 9,00 | 7,07 |
| 7000 | 27,88 | 48,50 | 26,40 | 19,90 | 12,80 | 9,03 | 7,10 |

| $\Sigma a_i^2$ | $S$ при $K=10$ | $L$ при различном числе выборок | | | | | |
|---|---|---|---|---|---|---|---|
| | | $K=5$ | $K=8$ | $K=10$ | $K=15$ | $K=20$ | $K=25$ |
| 7050 | 27,98 | 48,68 | 26,49 | 19,97 | 12,85 | 9,06 | 7,12 |
| 7100 | 28,08 | 48,86 | 26,58 | 20,04 | 12,89 | 9,10 | 7,15 |
| 7150 | 28,18 | 49,04 | 26,67 | 20,11 | 12,94 | 9,14 | 7,18 |
| 7200 | 28,28 | 49,22 | 26,76 | 20,18 | 12,99 | 9,17 | 7,20 |
| 7250 | 28,38 | 49,40 | 26,85 | 20,25 | 13,03 | 9,20 | 7,22 |
| 7300 | 28,48 | 49,58 | 26,94 | 20,32 | 13,08 | 9,23 | 7,25 |
| 7350 | 28,58 | 49,76 | 27,03 | 20,39 | 13,12 | 9,26 | 7,27 |
| 7400 | 28,68 | 49,94 | 27,12 | 20,46 | 13,17 | 9,29 | 7,29 |
| 7450 | 28,78 | 50,12 | 27,21 | 20,53 | 13,21 | 9,32 | 7,32 |
| 7500 | 28,85 | 50,30 | 27,30 | 20,60 | 13,25 | 9,35 | 7,35 |
| 7550 | 29,00 | 50,46 | 27,39 | 20,67 | 13,30 | 9,38 | 7,37 |
| 7600 | 29,12 | 50,62 | 27,48 | 20,74 | 13,34 | 9,41 | 7,40 |
| 7650 | 29,20 | 50,78 | 27,57 | 20,81 | 13,39 | 9,44 | 7,43 |
| 7700 | 29,27 | 50,94 | 27,66 | 20,88 | 13,43 | 9,47 | 7,45 |
| 7750 | 29,35 | 51,10 | 27,75 | 20,95 | 13,48 | 9,50 | 7,47 |
| 7800 | 29,44 | 51,26 | 27,84 | 21,02 | 13,52 | 9,53 | 7,50 |
| 7850 | 29,50 | 51,42 | 27,93 | 21,09 | 13,57 | 9,56 | 7,52 |
| 7900 | 29,60 | 51,58 | 28,02 | 21,16 | 13,61 | 9,59 | 7,54 |
| 7950 | 29,67 | 51,74 | 28,11 | 21,23 | 13,65 | 9,62 | 7,57 |
| 8000 | 29,78 | 51,90 | 28,20 | 21,30 | 13,70 | 9,65 | 7,60 |
| 8050 | 29,86 | 52,06 | 28,28 | 21,37 | 13,74 | 9,68 | 7,63 |
| 8100 | 29,98 | 52,22 | 28,36 | 21,44 | 13,78 | 9,71 | 7,66 |
| 8150 | 30,08 | 52,38 | 28,44 | 21,51 | 13,82 | 9,74 | 7,69 |
| 8200 | 30,18 | 52,54 | 28,52 | 21,58 | 13,86 | 9,77 | 7,71 |
| 8250 | 30,28 | 52,70 | 28,60 | 21,65 | 13,90 | 9,80 | 7,73 |
| 8300 | 30,37 | 52,86 | 28,68 | 21,72 | 13,94 | 9,83 | 7,76 |
| 8350 | 30,46 | 53,02 | 28,76 | 21,79 | 13,98 | 9,86 | 7,78 |
| 8400 | 30,55 | 53,18 | 28,84 | 21,86 | 14,02 | 9,89 | 7,80 |
| 8450 | 30,64 | 53,34 | 28,92 | 21,93 | 14,06 | 9,92 | 7,82 |
| 8500 | 30,73 | 53,50 | 29,00 | 22,00 | 14,10 | 9,95 | 7,84 |
| 8550 | 30,82 | 53,65 | 29,09 | 22,06 | 14,14 | 9,98 | 7,86 |
| 8600 | 30,91 | 53,80 | 29,18 | 22,12 | 14,18 | 10,01 | 7,89 |
| 8650 | 31,00 | 53,95 | 29,27 | 22,18 | 14,22 | 10,04 | 7,91 |
| 8700 | 31,09 | 54,10 | 29,36 | 22,24 | 14,26 | 10,07 | 7,93 |
| 8750 | 31,18 | 54,25 | 29,45 | 22,30 | 14,30 | 10,10 | 7,96 |



| $\Sigma a_i^2$ | $S$ при $K=10$ | $L$ при различном числе выборок | | | | | |
|---|---|---|---|---|---|---|---|
| | | $K=5$ | $K=8$ | $K=10$ | $K=15$ | $K=20$ | $K=25$ |
| 8800 | 31,27 | 54,40 | 29,54 | 22,36 | 14,34 | 10,13 | 7,98 |
| 8850 | 31,36 | 54,55 | 29,63 | 22,42 | 14,38 | 10,16 | 8,00 |
| 8900 | 31,45 | 54,70 | 29,72 | 22,48 | 14,42 | 10,19 | 8,02 |
| 8950 | 31,54 | 54,85 | 29,81 | 22,54 | 14,46 | 10,22 | 8,05 |
| 9000 | 31,63 | 55,00 | 29,90 | 22,60 | 14,50 | 10,25 | 8,07 |
| 9050 | 31,72 | 55,15 | 29,98 | 22,66 | 14,54 | 10,28 | 8,09 |
| 9100 | 31,81 | 55,30 | 30,06 | 22,72 | 14,58 | 10,31 | 8,11 |
| 9150 | 31,89 | 55,45 | 30,14 | 22,78 | 14,62 | 10,34 | 8,14 |
| 9200 | 31,98 | 55,60 | 30,22 | 22,84 | 14,66 | 10,37 | 8,16 |
| 9250 | 32,06 | 55,75 | 30,30 | 22,90 | 14,70 | 10,40 | 8,18 |
| 9300 | 32,15 | 55,90 | 30,38 | 22,96 | 14,74 | 10,43 | 8,20 |
| 9350 | 32,23 | 56,05 | 30,46 | 23,02 | 14,78 | 10,46 | 8,22 |
| 9400 | 32,32 | 56,20 | 30,54 | 23,08 | 14,82 | 10,49 | 8,25 |
| 9450 | 32,40 | 56,35 | 30,62 | 23,14 | 14,86 | 10,52 | 8,27 |
| 9500 | 32,49 | 56,50 | 30,70 | 23,20 | 14,90 | 10,55 | 8,29 |
| 9550 | 32,57 | 56,65 | 30,78 | 23,26 | 14,94 | 10,58 | 8,31 |
| 9600 | 32,65 | 56,80 | 30,86 | 23,32 | 14,98 | 10,60 | 8,34 |
| 9650 | 32,74 | 56,95 | 30,94 | 23,38 | 15,02 | 10,62 | 8,36 |
| 9700 | 32,82 | 57,10 | 31,02 | 23,44 | 15,06 | 10,65 | 8,38 |
| 9750 | 32,91 | 57,25 | 31,10 | 23,50 | 15,10 | 10,67 | 8,40 |
| 9800 | 32,99 | 57,40 | 31,18 | 23,56 | 15,14 | 10,70 | 8,42 |
| 9850 | 33,08 | 57,55 | 31,26 | 23,62 | 15,18 | 10,72 | 8,44 |
| 9900 | 33,16 | 57,70 | 31,34 | 23,68 | 15,22 | 10,75 | 8,46 |
| 9950 | 33,25 | 57,85 | 31,42 | 23,74 | 15,26 | 10,77 | 8,48 |
| 10000 | 33,33 | 58,00 | 31,50 | 23,80 | 15,30 | 10,80 | 8,50 |

Примечания.

1. Величины ошибок определены при вероятности их большего значения в 5%.

2 Таблица рассчитана на основе известных формул математической статистики; конкретные величины определялись при помощи логарифмической линейки, поэтому последняя цифра (сотые доли) не всегда надежна.

3. Лингвисту нет необходимости определять ошибку с точно-

стью до сотых; поэтому нужно прибегать к округлению табличных величин до десятых и единиц.

4. Если полученная в опыте сумма квадратов отклонений частот от их средней не дана в таблице, нужно искать наиболее близкую полученной в опыте табличную величину и внести (приближенно) поправку в табличную величину ошибки.

5. Величина несмещенной оценки среднего квадратичного отклонения дана в таблице лишь для 10 выборок; если опыт осуществлялся при ином числе выборок, нужно табличную величину несмещенной оценки среднего квадратичного отклонения умножить: при пяти выборках на 1,5; при восьми выборках на 1,1; при пятнадцати выборках на 0,8; при двадцати выборках на 0,7; при двадцати пяти выборках на 0,6.

### Как пользоваться таблицей.

Предположим, что десять выборок показали частоты, отклонения которых от средней, возведенные в квадрат, дали в сумме 4880; средняя частота — 90. Нужно найти ошибку наблюдения и определить интервал, в котором может лежать «действительная средняя».

В таблице нет суммы квадратов отклонений точно такой, какую нам дал опыт. Но там есть близкая к опытной сумма — 4900 — и ей соответствует, для 10 выборок, ошибка — 16,68; табличная сумма 4900 превышает полученную в опыте; но в таблице есть также сумма, несколько меньшая, чем опытная, — это сумма 4850, ей соответствует ошибка — 16,60; значит, ошибка, соответствующая опытной сумме квадратов отклонений частот от их средней, находится между 16,68 и 16,60, ближе к 16,68 (так как опытная сумма квадратов 4880 ближе к табличной сумме 4900, нежели к табличной же сумме 4850); это позволяет нам говорить, что показанная опытом ошибка наблюдения (выборки) равна приблизительно 16,66; но, видимо, указывать сотые доли для определения языковых частот не обязательно, поэтому можно принять, что наша ошибка приближенно равна 16,7; если же и десятые доли указывать не обязательно, ошибка будет приближенно равна 17 единицам.

3. Знание суммы квадратов отклонений выборочных частот от их средней и ее связи с величиной «хи-квадрат» позволяет построить таблицу, которая даст возможность по величине суммы квадратов отклонений, минуя вычисление величины «хи-квадрат», судить о превышении им границ существенности — 10%-ной, 5%-ной и 1%-ной, тем самым таблица разрешает лингвисту без вычисления величины «хи-квадрат» опираться на его показания в проверке нулевых гипотез.



Таблица № 2

**Величины $\sum a_i^2$, соответствующие числовым значениям $\chi^2$ для трех границ существенности**

| $\overline{x}$ | $K=5$ | | | $K=10$ | | |
|---|---|---|---|---|---|---|
| | 10% | 5% | 1% | 10% | 5% | 1% |
| 1 | 7,8 | 9,5 | 13,3 | 14,6 | 16,9 | 21,7 |
| 2 | 15,6 | 19,0 | 26,6 | 29,3 | 33,8 | 43,3 |
| 3 | 23,4 | 28,5 | 39,9 | 44,1 | 50,7 | 65,0 |
| 4 | 31,1 | 38,0 | 53,2 | 58,6 | 67,6 | 86,7 |
| 5 | 38,9 | 47,5 | 66,4 | 73,4 | 84,5 | 108,3 |
| 6 | 46,8 | 57,0 | 79,8 | 88,1 | 101,4 | 130,0 |
| 7 | 54,5 | 66,5 | 93,1 | 102,7 | 118,3 | 151,4 |
| 8 | 62,2 | 76,0 | 106,4 | 117,2 | 135,2 | 173,2 |
| 9 | 70,1 | 85,5 | 119,7 | 132,0 | 152,1 | 194,9 |
| 10 | 77,8 | 94,9 | 132,8 | 146,8 | 169,0 | 216,7 |
| 15 | 117 | 142 | 199 | 220 | 253 | 325 |
| 20 | 156 | 190 | 266 | 293 | 338 | 433 |
| 25 | 194 | 237 | 332 | 366 | 422 | 541 |
| 30 | 234 | 285 | 399 | 441 | 507 | 650 |
| 35 | 273 | 332 | 465 | 514 | 591 | 758 |
| 40 | 312 | 380 | 531 | 586 | 676 | 866 |
| 45 | 351 | 427 | 597 | 659 | 760 | 974 |
| 50 | 389 | 475 | 664 | 734 | 845 | 1083 |
| 55 | 429 | 522 | 730 | 807 | 929 | 1191 |
| 60 | 468 | 570 | 798 | 881 | 1014 | 1300 |
| 65 | 507 | 617 | 864 | 954 | 1098 | 1408 |
| 70 | 546 | 665 | 931 | 1027 | 1183 | 1514 |
| 75 | 585 | 712 | 997 | 1100 | 1267 | 1622 |
| 80 | 624 | 760 | 1062 | 1172 | 1352 | 1732 |
| 85 | 663 | 807 | 1128 | 1245 | 1436 | 1840 |
| 90 | 702 | 855 | 1197 | 1320 | 1521 | 1949 |
| 95 | 741 | 902 | 1263 | 1393 | 1605 | 2057 |
| 100 | 778 | 949 | 1328 | 1468 | 1688 | 2167 |
| 110 | 858 | 1044 | 1396 | 1615 | 1857 | 2384 |
| 120 | 936 | 1139 | 1529 | 1762 | 2026 | 2600 |
| 130 | 1014 | 1234 | 1727 | 1909 | 2195 | 2817 |
| 140 | 1092 | 1329 | 1859 | 2056 | 2364 | 3034 |

| $\bar{x}$ | $K=5$ | | | $K=10$ | | |
|---|---|---|---|---|---|---|
| | 10% | 5% | 1% | 10% | 5% | 1% |
| 150 | 1170 | 1424 | 1992 | 2202 | 2533 | 3250 |
| 160 | 1248 | 1519 | 2126 | 2349 | 2702 | 3467 |
| 170 | 1326 | 1614 | 2259 | 2495 | 2871 | 3681 |
| 180 | 1404 | 1709 | 2390 | 2642 | 3040 | 3899 |
| 190 | 1480 | 1804 | 2523 | 2788 | 3209 | 4116 |
| 200 | 1556 | 1898 | 2656 | 2935 | 3376 | 4334 |
| 210 | 1638 | 1993 | 2689 | 3082 | 3525 | 4551 |
| 220 | 1716 | 2088 | 2832 | 3228 | 3714 | 4767 |
| 230 | 1794 | 2183 | 2965 | 3376 | 3883 | 4984 |
| 240 | 1872 | 2378 | 3187 | 3521 | 4052 | 5200 |
| 250 | 1950 | 2473 | 3320 | 3669 | 4221 | 5397 |
| 260 | 2024 | 2468 | 3454 | 3816 | 4391 | 5634 |
| 270 | 2102 | 2563 | 3587 | 3962 | 4559 | 5848 |
| 280 | 2184 | 2658 | 3718 | 4112 | 4728 | 6068 |
| 290 | 2257 | 2753 | 3853 | 4255 | 4897 | 6283 |

Этой таблицей можно пользоваться для предварительного, ориентировочного решения вопроса о случайности или существенности наблюдавшегося в статистическом опыте расхождения (колебания) частот. Если сумма отклонений от средней, возведенных в квадрат, не превышает указанную в таблице величину для 5%-ного уровня существенности и полученной в опыте средней частоты (соответственно при пяти или десяти выборках),— нулевую гипотезу о случайности расхождения (колебания) частот можно принять; в противном случае, особенно если сумма квадратов отклонений значительно превышает табличную величину, гипотезу следует отвергнуть.

Можно сказать, что в таблице средние частоты, начиная с 10, даны с интервалами. Как быть, если опытная средняя частота окажется в одном из этих интервалов? Нужно представить полученную частоту как сумму двух частот, для которых суммы квадратов указаны, и сложить эти суммы. Например, при пяти выборках из текста получена средняя частота 155, и сумма отклонений от нее, возведенных в квадрат,— 1450. Можно ли принять гипотезу о случайности колебания наблюдавшихся частот? Нашу среднюю частоту можно представить как полученную из сложения 150 и 5; первой частоте в таблице соответствует



(на 5%-ном уровне существенности) число 1424; второй частоте (т. е. пяти) соответствует число 47,5; сложив их, получим 1471,5. Остается сравнить с этим табличным числом полученную в опыте сумму квадратов отклонений от средней (1450),— она несколько меньше табличной величины. Следовательно, гипотезу можно принять. Применение этой таблицы дает, хотя и не очень заметную, экономию времени наблюдателя языковых явлений.

4. Подобная же таблица может быть вычислена для сравнения двух выборочных средних по формуле $t = \frac{\bar{x}_1 - \bar{x}_2}{s_{1,2}} \sqrt{\frac{\kappa_1 \cdot \kappa_2}{\kappa_1 + \kappa_2}}$. В эту таблицу войдут полученные в опыте разности двух средних и соответствующие им минимальные на 5%-ном уровне существенности суммы $\Sigma a_{i1}^2 + \Sigma a_{i2}^2$. Такая таблица избавляет экспериментатора от многих вычислений.

Можно получить коэффициенты пересчета табличных значений $\Sigma a_{i1}^2 + \Sigma a_{i2}^2$ для случаев, когда $\kappa_1$ и $\kappa_2$ не равны десяти.

Вот некоторые из таких коэффициентов: а) при $\kappa_1 = \kappa_2 = 5$ коэффициент имеет величину 0,43; б) при $\kappa_1 = 10$ и $\kappa_2 = 5$ коэффициент — 0,67; в) при $\kappa_1 = 10$ и $\kappa_2 = 8$ коэффициент пересчета — 0,78; г) при $\kappa_1 = \kappa_2 = 15$ коэффициент — 2,56; д) при $\kappa_1 = \kappa_2 = 20$ коэффициент — 4,6.

Две иллюстрации применения таблицы.

Предположим, что две серии выборок из двух текстов дали нам средние частоты 195 и 175 (все выборки были равными) при $\Sigma a_{i1}^2 + \Sigma a_{i2}^2 = 5250$. Можно ли принять нулевую гипотезу о несущественности разности средних частот? Эта разность равна двадцати (195—175 = 20), в таблице такой разности соответствует величина 8200, опыт же дал нам 5250; но если минимальное значение — 8200, а мы получили 5250, нулевую гипотезу надо отклонить, разность средних существенна.

Предположим, что те же данные были получены в опыте, который состоял из 10 выборок одной серии и из 5 выборок другой; все выборки равного объема. В этом случае табличное значение $\Sigma a_{i1}^2 + \Sigma a_{i2}^2$ (т. е. 8200) нужно умножить на коэффициент 0,67; мы получим 5494. Выборочная величина и в этом случае меньше табличной, минимальной (5250 меньше 5494), поэтому гипотезу о несущественности разности средних частот нужно отклонить.

Таблица № 3

**Минимальные значения $\Sigma a_{i1}^2 + \Sigma a_{i2}^2$, соответствующие несущественной разности $\overline{x}_1$ и $\overline{x}_2$ (при $K_1 = K_2 = 10$ и $P = 5\%$)**

| $\overline{x}_1 - \overline{x}_2$ | $\Sigma a_{i1}^2 + \Sigma a_{i2}^2$ | $\overline{x}_1 - \overline{x}_2$ | $\Sigma a_{i1}^2 + \Sigma a_{i2}^2$ |
|---|---|---|---|
| 1,0 | 20,5 | 16,0 | 5248 |
| 1,5 | 46 | 16,5 | 5581 |
| 2,0 | 82 | 17,0 | 5925 |
| 2,5 | 128 | 17,5 | 6228 |

Таблица № 4

**Значения $K_\phi$, деление которого на $s_{1,2}$ дает выборочное значение $t$ для сравнения двух средних частот**

| $\overline{x}_1 - \overline{x}_2$ | $K_\phi$ | | | | |
|---|---|---|---|---|---|
| | $K_1 = K_2 = 5$ | $K_1 = 10$ $K_2 = 5$ | $K_1 = K_2 = 10$ | $K_1 = 10$ $K_2 = 15$ | $K_1 = K_2 = 15$ |
| 1,0 | 1,69 | 1,83 | 2,24 | 2,45 | 2,74 |
| 1,5 | 2,4 | 2,7 | 3,4 | 3,7 | 4,1 |
| 2,0 | 3,2 | 3,7 | 4,5 | 4,9 | 5,5 |
| 2,5 | 4,0 | 4,6 | 5,6 | 6,1 | 6,8 |
| 3,0 | 4,8 | 5,5 | 6,7 | 7,3 | 8,2 |

Такая таблица не избавляет от необходимости вычислять $s_{1,2}$, затрачивать на это некоторое время; и все же она заметно сокращает потери времени при использовании формулы Стьюдента для сравнения средних частот.

Пользование этой таблицей можно облегчить, если для вычисления $s_{1,2}$ использовать данные таблицы № 1, внося в них поправку с помощью особых коэффициентов пересчета.

Задача. Серия наблюдений из пяти выборок дала среднюю частоту 95 и сумму квадратов отклонений от нее 725; вторая серия наблюдений из 10 выборок дала среднюю частоту 75 и сумму квадратов отклонений от нее 1455. Можно ли расхождения средних признать существенными?

Объединение сумм квадратов отклонений дает 2180. В таблице № 1 этому значению соответствует величина, равная 15,55; ее нужно умножить на 0,84 (коэффициент пересчета, указанный в нижней строке таблицы № 4 — для случая, когда $K_1 = 5$ и $K_2 = 10$); умножение даст 13,2; это и есть величина $s_{1,2}$, на которую нужно разделить $K_\phi$, указанный в таблице № 4; так как выборочная разность средних равна 20, то значение $K_\phi$ равно 36,6. Делим 36,6 на 13,2, получаем выборочное значение $t$, равное 2,78; сравнив это значение с табличным, теоретическим (см. «Извлечение из таблицы числовых значений $t$») для 13 степеней свободы, так как в опыте было $K_1 + K_2 - 2 = 13$, и 5% уровня существенности, найдем, что выборочное значение $t$ больше табличного, поэтому расхождение средних частот нужно признать существенным.

Как применять таблицу № 4, если разность средних частот оказывается между значениями разности, указанными в таблице? Нужно полученную в опыте разность представить как сумму двух табличных значений разности. Например, если выборочная разность оказалась равной 25, находим по таблице $K_\phi$, соответствующий 20, и $K_\phi$, соответствующий 5, и складываем табличные величины.

Таблица № 4 несколько сложнее для применения, нежели таблица № 3, но зато дает более точные данные о существенности или случайности расхождения средних частот, позволяет устанавливать надежность принимаемых или отвергаемых гипотез, так как дает величину $t$.



6. Формула $\varepsilon_{1,2} = \sqrt{\bar{p} \cdot \bar{q} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, применяемая для сравнения двух выборочных долей, требует довольно больших вычислений. Для их облегчения и сокращения может быть использована таблица, рассчитанная автором для 32 различных объемов выборки (при $n_1 = n_2$) и для различных значений средней доли $p$.

Для того чтобы получить с помощью этой таблицы значение $\varepsilon_{1,2}$, надо перемножить табличные величины, соответствующие длине выборок и средней доле. Пусть из текста были взяты две выборки по 25 000 слов; доля некоторого явления $A$ в первой выборке оказалась равной 0,04, а во второй 0,08. Очевидно, среднюю долю мы получим, суммировав две выборочные и разделив сумму пополам: $\bar{p} = 0{,}06$. В таблице выборке длиной 25 000 соответствует величина 0,0089, а средней доле 0,06 — величина 0,238; перемножив эти величины, мы и получим $\Sigma_{1,2}$, равную 0,002. Теперь остается утроить полученное значение $\varepsilon_{1,2}$ и сравнить его с разностью долей: $\varepsilon_{1,2} \cdot 3 = 0{,}002 \cdot 3 = 0{,}006$, а разность долей $0{,}08 - 0{,}04 = 0{,}04$. Очевидно, таким образом, что утроенное квадратичное отклонение разности двух долей заметно меньше самой разности; это не позволяет сохранить нулевую гипотезу о несущественности расхождения двух долей.

Применение этой таблицы заметно сокращает и упрощает вычисления. Если лингвист встретится в опыте с объемами выборок и долями, отсутствующими в таблице, можно внести в табличные величины $\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ и $\sqrt{\bar{p} \cdot \bar{q}}$ поправки, приближенно прикинув различие между текстовыми выборками или долями и ближайшими выборками или долями в таблице. Особо строгая точность лингвисту не потребуется.

Итак, вниманию читателей были предложены пять таблиц, предназначенных для сокращения и упрощения различных вычислений, необходимых для применения статистических инструментов измерения частот, средних частот и долей. Возможно, профессионалы-математики не будут чересчур строгими к этим таблицам, так как составлены они не математиком и предназначены не для математиков.

Думается, такое назначение таблиц как-то объясняет их недостаточную, может быть, строгость и точность.

Таблица № 5

**Величины $\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ и $\sqrt{\overline{p}\cdot\overline{q}}$ в зависимости от числовых значений $n_1=n_2$ и $p$**

| $n_1=n_2$ | $\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ | $\overline{p}\,(\overline{q})$ | $\sqrt{\overline{p}\cdot\overline{q}}$ |
|---|---|---|---|
| 50 | 0,2000 | 0,01 (0,99) | 0,099 |
| 100 | 0,1414 | 0,02 (0,98) | 0,140 |
| 250 | 0,894 | 0,03 (0,97) | 0,171 |
| 500 | 0,0632 | 0,04 (0,96) | 0,196 |
| 750 | 0,0500 | 0,05 (0,95) | 0,218 |
| 1000 | 0,0447 | 0,06 (0,94) | 0,238 |
| 1250 | 0,0400 | 0,07 (0,93) | 0,255 |
| 1500 | 0,0365 | 0,08 (0,92) | 0,271 |
| 1750 | 0,0346 | 0,09 (0,91) | 0,286 |
| 2000 | 0,0316 | 0,10 (0,90) | 0,300 |
| 2500 | 0,0283 | 0,11 (0,89) | 0,313 |
| 3000 | 0,0265 | 0,12 (0,88) | 0,325 |
| 3500 | 0,0239 | 0,13 (0,87) | 0,336 |
| 4000 | 0,0224 | 0,14 (0,86) | 0,347 |
| 4500 | 0,0210 | 0,15 (0,85) | 0,357 |
| 5000 | 0,0200 | 0,16 (0,84) | 0,367 |
| 7000 | 0,0170 | 0,18 (0,82) | 0,384 |
| 8500 | 0,0152 | 0,19 (0,81) | 0,392 |
| 10000 | 0,0141 | 0,20 (0,80) | 0,400 |
| 15000 | 0,0114 | 0,22 (0,78) | 0,414 |
| 20000 | 0,0109 | 0,25 (0,75) | 0,433 |
| 25000 | 0,0089 | 0,27 (0,73) | 0,444 |
| 40000 | 0,0071 | 0,30 (0,70) | 0,458 |
| 50000 | 0,0063 | 0,32 (0,68) | 0,466 |
| 75000 | 0,0049 | 0,35 (0,65) | 0,477 |
| 100000 | 0,0045 | 0,37 (0,63) | 0,483 |
| 200000 | 0,0032 | 0,40 (0,60) | 0,490 |
| 400000 | 0,0022 | 0,42 (0,58) | 0,494 |
| 500000 | 0,0020 | 0,45 (0,55) | 0,497 |
| 750000 | 0,0017 | 0,47 (0,53) | 0,499 |
| 1000000 | 0,0014 | 0,50 (0,50) | 0,500 |



## УЧЕНИЕ О СТИЛЯХ ЯЗЫКА И СТИЛЯХ РЕЧИ И СТАТИСТИКА

1. Едва ли в настоящее время может быть дискуссионным вопрос о применимости статистической методики в изучении языка,— этот вопрос решен положительно практикой лингвистических исследований. Выполнен не один десяток больших и малых работ, посвященных статистическому изучению русского и других языков. В задачи автора не входит анализ или хотя бы обзор этих работ. Известное представление о них можно получить по краткому библиографическому указателю, приложенному к этой книге.

Но все же хочется упомянуть о том, что статистика испытана и дала положительные результаты и в области фонетики, и в области лексики, и в области морфологии и синтаксиса, и в области языковых и речевых стилей. Совсем недавно лингвисты получили две серьезные и обширные по обследованному материалу работы — «Статистико-комбинаторные методы в теоретическом и прикладном языкознании» Н. Д. Андреева и «Статистические параметры стилей» группы молодых украинских лингвистов, возглавляемой В. И. Перебийнос. Можно согласиться или не соглашаться с отдельными сторонами описанных в этих работах экспериментов, можно спорить с отдельными теоретическими решениями, но нельзя не видеть того, что статистика все смелее и шире применяется учеными для осмысления сложнейших проблем лингвистической науки, в их числе — проблемы языковых и речевых стилей.

Как известно, теория языковых и речевых стилей складывается медленно и трудно. О противоречиях и исканиях в этой области науки дает некоторое представление дискуссия, шедшая на страницах журнала «Вопросы языкознания» в 1954 г.; ее результаты были обобщены В. В. Виноградовым [1].

Дискуссия и опубликованные после нее работы показали, как велик диапазон колебаний лингвистической мысли, пытающейся схватить сущность стилей, и как далека еще лингвистика от достижения хотя бы неполного единства в оценке основных понятий учения о стилях.

[1] См.: В. В. Виноградов. Итоги обсуждения вопросов стилистики. «Вопросы языкознания», 1955, № 1.

После дискуссии появилось несколько интересных работ, посвященных проблемам стилистики[1]. Среди них особое место занимает книга В. В. Виноградова «Стилистика. Теория поэтической речи. Поэтика». Автор убедительно обосновывает необходимость различения, по крайней мере, трех стилистик — стилистики языка, стилистики речи, стилистики художественной литературы.

Изучению и более глубокому пониманию стилей помогают статистические данные, полученные разными лицами и опубликованные лишь частично[2].

Все отчетливее вырисовывается необходимость обогащения привычных (хотя все еще весьма нечетких) качественных представлений о стилях речи представлениями количественными. Уже говорилось о том, что возможность и необходимость применения количественных оценок языка и речи заложены в самой природе функционирования и развития языка. Соответствующие иллюстративные примеры читателю были предложены. Количество таких примеров легко может быть умножено.

Конечно, далеко не всегда статистический опыт дает очевидные, ясно выраженные доказательства действия некоторого статистического закона. Нередки случаи, когда статистический закон испытывает нарушающие его чистоту воздействия. Но это не удивительно: единицы и категории языка в своем функционировании и развитии подчинены многим и разным влияниям, совокупность и система которых колеблется, вызывая перебои в действии статистического закона, как бы наложение одной вероятности на другую. Удивляться приходится не нарушением статистических законов в конкретных текстах, а тому, как эти законы в общем хорошо сохраняются работающей структурой языка, что и позволяет убежденно применять статистическую методику в лингвистических исследованиях.

К тому же полезно вспомнить и о том, что изменение частот некоторых изучаемых элементов языка меняет качество речевых структур. Так, структура *АВСАДБЕАЕЕЕ* качественно отличается от структуры *ВВАСВВДДДЕД*, хотя состав элементов в той и другой структуре одинаков (*А, В, С, Д, Е*). Конечно, изменение качества речевых структур зависит не только от перемен в количественных соотношениях одних и тех же элементов, но для нас сейчас важно то, что и от таких перемен качество речевых структур несомненно зависит. Этот факт существенно важен для обоснования количественного подхода к стилям языка и стилям речи, для формирования обновленного, вероятностно-статистического понимания тех и других.

2. В 1963 г. Н. Д. Андреев и Л. Р. Зиндер вводят в научный обиход понятие речевой вероятности. Они пишут: «Если обратиться к парадигмам русского словоизменения, то шесть падежей представляются равноправными в системе языка. Поведение же их в синтагматике речи свидетельствует о том, что здесь между ними нет никакого равноправия: именительный и родительный встречаются в несколько раз чаще творительного и во много раз чаще дательного падежа. Рассмотрение подобного рода фактов, заимствованных из речи, свидетельствует о том, что некоторые ее характеристики, притом далеко не маловажные, не могут быть выделены из системы языка. Мы приходим таким путем к существованию еще и другого отвлечения от речи, которое уместно назвать речевой вероятностью.

Система речевой вероятности в таком понимании есть совокупность относительных количественных характеристик, описывающих численные соотношения между элементами (или группами элементов) в некотором массиве текстов. Можно сказать, что речевая вероятность определяет статистическую структуру текстов, тогда как язык характеризуется их теоретико-множественной структурой и алгоритмами их порождения и распознавания. Таким

---

[1] См.: например: В. В. Виноградов. Стилистика. Теория поэтической речи. Поэтика. М., 1963; В. Д. Левин. Очерк стилистики русского литературного языка конца XVIII — начала XIX в. М., 1964; М. Н. Кожина. О специфике художественной и научной речи в аспекте функциональной стилистики. Пермь, 1966; Р. Г. Пиотровский. Очерки по стилистике французского языка, Л., 1960, Ю. С. Степанов. Французская стилистика. М., 1965, и др.

[2] См. работы Н. Д. Андреева, Г. Г. Белоногова, С. И. Кауфмана, А. Г. Лескиса, В. А. Никонова, коллектива под руководством В. И. Перебийнос, Р. М. Фрумкиной, коллектива саратовских ученых под руководством О. Б. Сиротининой; на кафедре русского языка и общего языкознания Горьковского университета выполнено или выполняется пять кандидатских диссертаций и более 75 дипломных работ, содержащих неопубликованный статистический материал, показывающий функциональные и индивидуальные стили русского литературного языка XIX—XX веков.

образом, целесообразно различать четыре понятия: речевой акт, речь, речевая вероятность и язык» [1].

Несомненно, что авторы нарисовали отнюдь не стандартную научную картину. Однако она требует некоторых уточнений, если ее внимательно сравнить с оригиналом, с которого и для понимания которого она исполнена.

Но для того чтобы желаемые уточнения сделать по возможности корректно, обратимся еще раз к мыслям Н. Д. Андреева, высказанным уже не в 1963, а в 1967 г.: «Если взять из лото 32 бочонка, расклеить на них русский алфавит и перемешать, то вероятность того, что первый же вынутый бочонок окажется с буквой, изображающей чистую гласную (так!— *Б. Г.*), будет определена дробью 6 : 32, т. е. будет несколько менее 19%. Если же взять произвольный русский текст и выбрать из него наугад одну букву, то вероятность того, что она окажется знаком чистой гласной (так! — *Б. Г.*). будет приближенно равна 30%, колеблясь вверх или вниз от этой величины в зависимости от типа текста.

При первом выборе, дающем 19%, мы будем иметь дело с парадигматической вероятностью нашей группы из 6 букв, при втором, дающем примерно 30%,— с их же синтагматической вероятностью (частотностью)».

«Парадигматические вероятности принадлежат языку и всегда равны у разных единиц одной группы; синтагматические вероятности принадлежат речи и, за редкими исключениями, не равны у различных единиц одной и той же группы» [2].

Но ведь, по-видимому, языковая (парадигматическая) вероятность (в понимании и описании Н. Д. Андреева) реальному, функционирующему и развивающемуся языку вообще неизвестна. Предположить, что все гласные звуки или все формы падежа, или все члены предложения равновероятны,— это значит подменить естественный живой язык его схемой. Но язык надо рисовать таким, каков он

[1] Н. Д. Андреев и Л. Р. Зиндер. О понятиях речевого акта, речи, речевой вероятности и языка. «Вопросы языкознания» 1963, № 3, стр. 18—19.

[2] Н. Д. Андреев. Статистико-комбинаторные методы в теоретическом и прикладном языкознании. Л., 1967, стр. 17.



есть. А это, в частности, означает, что единицы и категории языка нужно видеть в их реальном функционировании, в котором они реально же обладают вероятностями, как их объективными свойствами. И эти их вероятности меняются в зависимости от многих условий, но прежде всего в зависимости от специфических коммуникативных потребностей различных видов и типов человеческой деятельности, общественной жизни.

Вероятность — это объективное свойство функционирующего и развивающегося языка и, естественно, что об этом свойстве мы узнаем, наблюдая применение элементов языка в речевых актах коммуникации, в различных речевых структурах. Каждая языковая единица имеет и качественные, и количественные характеристики, свойства, признаки. И те и другие объективны. Гласные звуки русского языка отличаются друг от друга качественно — наборами своих дифференциальных признаков, своими соотношениями с другими гласными и согласными звуками и т. д. Но те же самые звуки характеризуются и своими вероятностями, своей готовностью к речевому коммуникативному применению, своей коммуникативной активностью, обнаруживаемой каждый раз в тех частотных соотношениях звуков, которые наблюдатель устанавливает в определенном, достаточно обширном массиве речи. По данным группы авторов, вероятность гласных фонем русского языка в современной речи такова: фонема |а| и ее варианты — 0,185, фонема |и| и ее варианты — 0,122, фонема |э| и ее варианты — 0,047, фонема |у| — 0,039, фонема |о| — 0,038, фонема |ы| — 0,019 [1]. Это значит, что звуковой механизм нашего языка работает таким образом, что на каждые 1000 звуков в речевом потоке «выдает» в среднем 185 звуков [а], 122 звука [и], 47 звуков [э], 39 звуков [у], 38 звуков [о] и всего 19 звуков [ы]. Хорошо видно, что речевая активность фонем |а| и |и| во много раз превосходит речевую активность фонемы |ы|.

Правда, сообщенные только что сведения о вероятностях русских гласных должны быть отнесены к работе фонемного механизма нашего языка без учета возможных стилевых вариантов его деятельности. Как правило, единицы и категории языка получают в его речевой деятель-

[1] См. об этом: М. А. Сапожков. Речевой сигнал в кибернетике и связи. М., 1963, стр. 62.

ности ту или иную дифференциацию. Поэтому, например, в нашей речи, в различных ее видах и типах, существуют и выявляются различные вероятности падежных форм, форм времени, предложений различной структуры и т. д. Так что нет вероятности родительного падежа вообще, а есть реальная вероятность родительного падежа в художественной прозе и публицистике, в науке и бытовой речи. Вероятности языковых единиц и категорий могут меняться в зависимости от области применения языка, целей его использования и т. д. Возникает поэтому теоретическая и прикладная проблема варьирования вероятностей одних и тех же языковых единиц, проблема стилевой дифференциации вероятностей и выявляющих эти вероятности частот.

3. В настоящее время стили языка обычно рассматриваются (если не принимать во внимание всего разнообразия индивидуальных мнений и взглядов) как его структурные разновидности, или как подсистемы его системы, или как области его структуры. Вот некоторые, очень немногие, даваемые лишь для иллюстрации, определения:

«Языковой стиль — это разновидность общенародного языка, сложившаяся исторически и характеризующаяся известной совокупностью языковых признаков, часть из которых своеобразно, по-своему повторяется в других языковых стилях, но определенное сочетание которых отличает один языковой стиль от другого» [1].

«С точки же зрения собственно языковой — стили речи — это осознанные обществом разновидности общенационального языка, закрепленные за типами поведения и деятельности людей в обществе и характеризующиеся: 1) отбором средств (слов, типов предложения, типов произношения) из общенационального языкового достояния и 2) скрытым за этими средствами общим принципом отбора» [2].

«...Разновидности и формы функционирования находятся в постоянном взаимодействии и взаимопересечении. На основе этого взаимодействия складываются так называемые стили общенародного языка (языковые стили), представляющие собой замкнутые, не пересекающиеся, общественно осознанные системы отбора языкового материала» [1].

Стиль — «совокупность приемов использования средств языка, характерная для какого-либо писателя или литературного произведения, направления, жанра... Совокупность особенностей в построении речи и словоупотреблении, манера словесного изложения» [2].

«Если исходить из понимания общей структуры языка как «системы систем» (что вызывает отдельные возражения, не всегда достаточно обоснованные), то стиль языка — это одна из частных систем (или «подсистем»), входящих в общую систему. Понятие «стиля языка» в основном определяется теорией функций языка в их реальном разграничении. В этом аспекте стиль языка — это структурный облик функции языка в ее многообразных проявлениях» [3].

Это всего лишь иллюстрации определившихся или намечающихся вновь подходов к пониманию стилей языка и стилей речи. Многое продолжает оставаться неясным, многое — спорным. Нет еще желаемой четкости в разграничении и соотнесении стилей языка и стилей речи, хотя правомерность самой проблемы теперь, в особенности после известных работ В. В. Виноградова, едва ли может оспариваться. Не ясно, нужно ли искать существенные признаки стиля в области самой языковой структуры или же, в первую очередь, в сферах ее функционирования. Понятия и термины «отбор» и «выбор», применяемые нередко в описаниях языковых стилей, кажутся некорректно предложенными читателю и неуместными при разъяснении сути вещей. Намечаемые как будто в некоторых суждениях о стилях количественные представления и характеристики очень неопределенны. Как устанавливать границы между стилями, неизвестно. Участие разных уровней языковой структуры в образовании стилей, в сущности, не изучено и линии такого изучения зыбки и неотчетливы. Одним словом, множество проблем и вопросов, связанных с изучением и пониманием стилей языка и стилей речи, ждет ответа лингвистов.

---

[1] Р. А. Будагов. Введение в науку о языке. М., 1965, стр. 467.
[2] Ю. С. Степанов. Основы языкознания, М., 1966, стр. 170.



[1] Р. Г. Пиотровский. Очерки по стилистике французского языка. Л., 1960, стр. 19.
[2] Словарь современного русского литературного языка, т. 14, М. — Л., 1963, стлб. 877.
[3] В. В. Виноградов. Стилистика. Теория поэтической речи. Поэтика, М., 1963, стр. 201.

Поэтому и возникает мысль: нельзя ли, не отказываясь от новых и новых попыток качественного (так сказать, традиционного) определения и истолкования языковых (и речевых) стилей, попытаться подойти к ним с позиций количественных представлений и характеристик и в соответствии с этим целенаправленно, объективно и систематически изучать стили при помощи статистической методики? Такая мысль теперь получает опору в опытах статистической оценки различных элементов стилей.

Применим модную в настоящее время методику моделирования. Представим себе, что некоторый условный язык имеет всего 10 слов и что эти слова мы обозначили буквами *А, В, С, Д, Е, К, Л, М, Н, О*. Язык функционирует, обслуживая потребности науки, деловой деятельности государства и общества, художественной литературы. Эти потребности различны, вследствие чего активность одних и тех же слов в науке, деловой деятельности и художественной литературе неодинакова. Так, обслуживая науку, наш условный язык обнаруживает следующие вероятности входящих в него слов: *А* — 0,02; *В* — 0,03; *С* — 0,05; *Д* — 0,25; *К* — 0,17; *Л* — 0,10; *М* — 0,01; *Н* — 0,02; *О* — 0,02; *Е* — 0,33. Обслуживая деловую деятельность государства и общества, наш язык имеет такие вероятности своих слов: *А* — 0,20; *В* — 0,25; *С* — 0,03; *Д* — 0,27; *Е* — 0,10; *К* — 0,02; *Л* — 0,03; *М* — 0,04; *Н* — 0,01; *О* — 0,05. В художественной литературе слова нашего языка обнаруживают такую вероятность: *А* — 0,09; *В* — 0,11; *С* — 0,10; *Д* — 0,05; *Е* — 0,08; *К* — 0,07; *Л* — 0,12; *М* — 0,18; *Н* — 0,11; *О* — 0,09.

Что это значит, если перейти от вероятностных абстракций к речевым цепочкам, построенным из слов нашего языка? А это значит, в частности, то, что такие речевые цепочки окажутся очень различными качественно в науке, деловом документе и в художественном произведении. В науке: *ДКЕЕДВКДЕ*... В деловых документах: *АВДДАВОЕДВ*... В художественной литературе: *АВСЕКЛМНМО*... (Конечно, порядок следования слов показанными вероятностями не предусмотрен, но в реальном языке и он подчинен статистическим закономерностям.)

Можно думать, что таким образом мы получаем удовлетворительную модель стилей языка. Она может быть значительно улучшена, если в нее ввести вероятности не только слов самих по себе, но и их морфологических классов, их синтаксических позиций, их сочетаемости, их порядка следования, их интонирования и т. д.— в соответствии с теми элементами и признаками реально работающей языковой структуры, которые нам известны.



Ведь и в реальном, естественном языке каждая его единица имеет вероятность своего применения (встречаемости) в потоке других единиц, вероятность своего позиционного использования, своей сочетаемости с другими единицами, своего участия в построении предложений и т. д.

Так намечается обновленный, вероятностно-статистический подход к пониманию стилей языка, позволяющий сформулировать и их обновленное определение: может быть, функциональные стили языка — это типы его функционирования, соответствующие различиям социальной практики коллектива и отличающиеся друг от друга существенными различиями вероятностей языковых единиц и их категорий, достаточными для их совокупного качественного опознавания людьми на интуитивном уровне восприятия речи.

В этом определении центр внимания перемещается с разновидностей языковой структуры на варианты (типы) ее функционирования. Если язык уподобить очень сложному механизму, способному работать в нескольких различных режимах, предполагающих неодинаковое участие отдельных узлов и составных частей, то стили языка допустимо было бы сравнить как раз с этими разными режимами работы. Так главной задачей исследователя стилей становится не поиск вариантов структуры или подсистем системы, а объективное изучение вероятностных характеристик всех языковых единиц и категорий на всех структурных уровнях. Такое изучение не отвергает, а предполагает и качественные оценки и определения, а в перспективе и более глубокое качественное понимание и описание стилей, но уже на прочной основе строго оцененных с помощью статистики фактов.

Опыты показывают, что интуитивные представления о пяти или шести главных функциональных стилях современного русского литературного языка подтверждаются данными статистики. Вместе с тем становится очевидной

неполнота и бедность привычных представлений о языковых стилях, в особенности об их структурных признаках и о линиях их различия и разграничения.

Вот некоторые предварительные иллюстративные данные о вероятностях отдельных элементов языка, вычисленные на основе сведений о частотах этих же элементов в речи [1].

**Вероятность применения действительного, страдательного и среднего залогов**

| Залоги | Стили | | |
|---|---|---|---|
| | художественный | научный | публицистический |
| Действительный | 0,42 | 0,47 | 0,55 |
| Страдательный | 0,01 | 0,09 | 0,07 |
| Средний | 0,56 | 0,43 | 0,37 |

Художественный стиль в опыте был представлен авторской речью Л. Толстого, А. Толстого и М. Шолохова; научный стиль — речью И. П. Павлова и А. Г. Столетова; публицистический — речью газетных статей: различия между авторами были сняты путем усреднения частот и, соответственно, вероятностей; в выборки входили только глаголы, и длина совокупной выборки по каждому стилю колебалась от 7500 до 15000 применений глагола; колебания вероятностей в таблицу не включены.

Таблица показывает, что, например, на каждые сто глаголов, встретившихся последовательно читателю в тексте художественном, приходится всего один случай применения глагола страдательного значения, а в тексте научном — девять случаев. Также в среднем на каждые сто употреблений глагола в художественном произведении читатель встретит 42 глагола действительного залога, а в газете — 55, т. е. значительно больше. В художественном произведении на каждые сто глаголов встретится в среднем 56 со среднезалоговыми признаками, а в статьях публициста значительно меньше — 37.

[1] Сведения о частотах получены студентами Горьковского университета И. Урамбашевым, М. Маблибовской, Л. Шапошниковой, Л. Трофимовой и сообщены в их дипломных работах.



**Вероятность применения частей речи**

| Части речи | Стили | | |
|---|---|---|---|
| | художественный | научный | публицистический |
| Имя существительное | 0,40 | 0,47 | 0,53 |
| Имя прилагательное | 0,15 | 0,23 | 0,22 |
| Имя числительное | 0,01 | 0,01 | 0,02 |
| Местоимение | 0,12 | 0,06 | 0,06 |
| Глагол | 0,18 | 0,09 | 0,13 |
| Причастие | 0,03 | 0,05 | 0,03 |
| Деепричастие | 0,005 | 0,003 | 0,005 |
| Наречие | 0,07 | 0,08 | 0,04 |
| Предлог | 0,16 | 0,16 | 0,12 |
| Союз | 0,11 | 0,08 | 0,06 |

Вероятности различных частей речи не подвергаются здесь (как и в первой таблице) статистической проверке, т. е. не дается ошибка их вычисления и не принимаются во внимание суммарные объемы выборок по каждому стилю. Можно все же сказать, что научный и публицистический стили представлены были в опыте суммарными выборками длиной в 10000 знаменательных слов (точнее — словоупотреблений); суммарная выборка из текстов художественных была равна приблизительно 75000 словоупотреблений и была получена группой студентов-дипломников, фамилии которых я уже называл в печати. В основе опыта лежали малые выборки одинаковой длины — в 500 знаменательных слов каждая. Поэтому показанная в таблице вероятность предлогов и союзов должна пониматься как вероятность по отношению к употреблению знаменательных слов, а не вообще всех слов текста.

Внимательно вглядываясь и вдумываясь в таблицу, нетрудно заметить, что имена существительные и имена прилагательные в художественном стиле заметно менее активны, чем в научном и публицистическом; что местоимения и глаголы, наоборот, заметно более активны в художественном по сравнению с научным и публицистическим; что предлоги менее активны в публицистическом стиле, нежели в художественном и научном; что союзы более

активны в художественном, чем в научном и публицистическом; что причастие не обнаруживает заметных симпатий ни к одному из стилей, то же — имя числительное, и т. д.

Уже этими скупыми данными подрывается привычное представление о стилевом нейтралитете морфологии, в особенности на ее самом абстрактном уровне распределения слов по частям речи. Остро встает задача широкого стилистического изучения стилевой дифференциации частей речи, их форм и категорий, всей морфологической структуры русского литературного языка в ее современном функционировании и историческом движении.

В пояснение сказанного приведем еще две таблицы.

**Вероятность применения отдельных падежей в речевом ряду имен существительных [1].**

| Падежи | Стили | |
|---|---|---|
| | художественный | публицистический |
| Именительный | 0,28 | 0,18 |
| Родительный | 0,17 | 0,36 |
| Родительный с предлогом | 0,05 | 0,04 |
| Дательный | 0,02 | 0,02 |
| Дательный с предлогом | 0,06 | 0,03 |
| Винительный | 0,13 | 0,13 |
| Винительный с предлогом | 0,06 | 0,05 |
| Творительный | 0,07 | 0,04 |
| Творительный с предлогом | 0,07 | 0,04 |
| Предложный | 0,08 | 0,10 |

Всего в художественном стиле 500 существительным подчиняется в среднем 361 слово, а в публицистическом стиле — 504 слова.

Таблицы убедительно говорят о том, что не только соотношения частей речи, но и их формы и категории, их сочетаемость в той или иной мере дифференцированы по стилям, т. е. участвуют в стилеобразовании. Достаточно сейчас обратить внимание хотя бы на то, как различно соотнесены именительный и родительный падежи в художественном и публицистическом стилях: в художественном именительный господствует над родительным, в публицистическом — родительный над именительным. Не менее интересны и данные о сочетаемости имени существительного: оказывается, в художественном стиле подчинение именам существительным глаголов заметно активнее подчинения других существительных; в публицистике же подчинение существительным других существительных явно господствует над подчинением существительным глаголов. А это, между прочим, ведет к тому, что словосочетания, построенные по схеме «имя существительное плюс имя существительное», очень активны в публицистике, что и является одной из структурных причин возникновения различных газетных штампов, от которых пытается защищаться художественная речь, испытывающая мощное воздействие речи публицистической.

[1] Вероятности вычислены на основе данных Ж. Трофимовой (требуют дальнейших уточнений).



**Вероятность подчинения в тексте имени существительному различных частей речи [1].**

| Что подчиняется | Стили | |
|---|---|---|
| | художественный | публицистический |
| Имя существительное | 0,18 | 0,35 |
| Имя прилагательное | 0,25 | 0,32 |
| Глагол | 0,26 | 0,14 |
| Причастие | 0,08 | 0,03 |
| Местоимение | 0,13 | 0,09 |
| Прочие | 0,10 | 0,07 |

В только что предложенных вниманию читателя таблицах сравниваются вероятности некоторых категорий русской морфологии в двух или трех стилях (оценка точности определения вероятностей не дана, чтобы не усложнять изложения). Читатель, естественно, замечает большие или меньшие расхождения вероятностей одной и той же категории в разных стилях. Если применить описанные ранее инструменты сравнения долей, то можно установить, что в одних случаях показанные таблицами расхождения вероятностей существенны, в других — случайны. Построим новую таблицу, в которой существенные

[1] Вероятности вычислены на основе данных Н. Лавровой (требуют дальнейших уточнений).

расхождения вероятностей некоторых морфологических категорий между художественным и публицистическим стилями отметим знаком плюс, а случайные — знаком минус.

| Категории | Стили | |
|---|---|---|
| | художественный | публицистический |
| Имя существительное | + | + |
| Имя прилагательное | + | + |
| Местоимение | + | + |
| Глагол | + | + |
| Причастие | — | — |
| Деепричастие | — | — |
| Наречие | + | + |
| Предлог | + | + |
| Союз | + | + |
| Действительный залог | + | + |
| Страдательный залог | + | + |
| Средний залог | + | + |
| Именительный падеж | + | + |
| Родительный падеж | + | + |
| Дательный падеж | — | — |
| Винительный падеж | — | — |
| Творительный падеж | + | + |
| Предложный падеж | + | + |
| Существительное подчиняется существительному | + | + |
| Прилагательное подчиняется существительному | + | + |
| *Глагол подчиняется* существительному | + | + |
| Местоимение подчиняется существительному | + | + |
| Причастие подчиняется существительному | + | + |

Таким образом, на основании этой новой таблицы можно говорить о том, что одни вероятности различают, дифференцируют стили языка (по крайней мере, те два стиля, которые отображены в таблице), другие же вероятности, хотя и могут различаться арифметически, но



это их различие случайно и потому сами такие вероятности не дифференцируют стилей языка. Первые вероятности назовем дифференцирующими стилевыми вероятностями, вторые — нейтральными стилевыми вероятностями. Первые существенно, вторые случайно отклоняются друг от друга.

Ничто, в принципе, не мешает получить статистические данные, характеризующие все морфологические и синтаксические (а затем и все лексические, словообразовательные, фонетические) явления языка на основе обследования разных видов и типов речи.

Это позволило бы построить принципиально новое описание языковых стилей: каждый из них определялся бы прежде всего наборами дифференцирующих и нейтральных вероятностей. При этом очевидно, что, чем обширнее наборы дифференцирующих вероятностей и чем больше сами такие вероятности, тем отчетливее противопоставлены и яснее осознаются языковые стили.

Здесь уместно сказать о том, что термин «нейтральные стилевые вероятности» несколько условен, потому что за ним стоят, в сущности, колебания одной и той же вероятности, ее варианты в различных типах и видах речи. (Вместо терминов «дифференцирующие стилевые вероятности» и «нейтральные стилевые вероятности» можно было бы предложить также термины «дифференцирующие стилевые доли» и «нейтральные стилевые доли»).

Уже осуществленные опыты говорят о том, что главные функциональные стили языка, выделенные лингвистикой без применения статистического аппарата и статистической методики, существуют и ждут объективного и полного описания. Вместе с тем опыты показывают, что главные функциональные стили имеют внутреннюю дифференциацию на жанровые, тематические и даже индивидуальные разновидности. Так, разновидность публицистического стиля, соотнесенная с жанром заметок и сообщений, может по ряду признаков существенно отличаться от разновидности, соотнесенной с жанром передовой статьи: вероятности какого-то круга грамматических категорий и явлений лексики окажутся в газетных передовицах и в газетных информационных сообщениях дифференцирующими, т. е. существенно различными. Разновидность на-

учного стиля, обслуживающая нужды теоретической физики, может оказаться отделенной некоторым набором дифференцирующих вероятностей (долей) от разновидности, обслуживающей нужды математики или биологии. Индивидуальная разновидность художественного стиля, формируемая и применяемая в прозе Л. Леонова, может существенно отличаться от индивидуальной разновидности того же стиля, формируемой и применяемой в прозе К. Паустовского, и отличие этих разновидностей друг от друга обязательно будет показано теми или иными наборами дифференцирующих стилевых вероятностей (долей).

Первоначальное представление о языковых стилях возникает из наблюдения видов и типов речи, различаемых интуитивно. Каждый вид или тип речи образует некоторую более или менее самостоятельную систему, порождаемую соответствующим языковым стилем. Однако из таких систем не всегда можно извлечь информацию о достаточном числе дифференцирующих стилевых вероятностей: нередко нейтральных стилевых вероятностей оказывается больше, чем дифференцирующих. Иначе говоря, строгой картины стилевых границ не получается. Их и не может, по-видимому, быть, так как очень сложны комбинации и переплетения многих внеязыковых и внутриязыковых условий, от совокупного действия которых зависит функционирование языка, варьирование его работы.

Применение вероятностно-статистической методики и соответствующих теоретических представлений должно помочь науке о языке объективнее, доказательнее и точнее описать основные функциональные стили языка и их варианты, установить более выраженные и менее выраженные реальные границы между ними.

5. Если единицы и категории языка объективным своим признаком (или — свойством) имеют, в числе других, вероятность, то элементы речевой последовательности, структура речи, обладают частотами, в которых находит себе выражение языковая вероятность.

Взаимосвязь вероятностей и частот сама по себе представляет интереснейшую проблему для лингвиста, применяющего статистическую методику. В плане общей теории языка эта взаимосвязь говорит о сложном единстве языка и речи, о том, что язык и речь — два вида социального бытия той коммуникативной структуры, которую человечество выработало и применяет как «действительное, практическое сознание», по определению К. Маркса. Эта лингвоструктура социально существует либо в виде языка, либо в отвлечении от какого бы то ни было конкретного логического или эмоционального, эстетического и иного содержания, либо в виде речи, т. е. в связи с таким содержанием. В отвлечении от конкретного содержания эта лингвоструктура представляет собой совокупность и систему единиц общения и их категорий, готовых к речевой реализации. Вступив в связь с конкретным содержанием, лингвоструктура представляет собой последовательность единиц общения, образующую особую структуру речи.

В плане вероятностно-статистического изучения языка и речи взаимосвязь вероятностей и частот интересна для лингвиста уже тем, что позволяет, наблюдая и статистически обрабатывая частоты, переходить от них к вероятностям (а значит, и статистическим законам), а зная вероятности, переходить от них к частотным характеристикам речевых структур.

Теперь хорошо известно, что очень многие единицы языка и их категории имеют устойчивые частоты. Соответствующие иллюстрации вниманию читателя уже были предложены. Эта устойчивость частот многих единиц языка в речи позволяет, на основе выборочного изучения, измерять большие речевые массивы, устанавливать свойственные таким массивам частоты и пучки частот.

Интуитивно, читательским восприятием различаемые типы, виды (или стили) речи и отличаются друг от друга, по-видимому, прежде всего устойчивыми (а возможно, также и неустойчивыми) частотами и их пучками. Определить, в каких случаях перед нами случайное расхождение частот, а в каких — существенное, позволяет математическая статистика, даже тот ее несложный аппарат, который уже был показан читателю и который минимально необходим каждому лингвисту. Опытов такого применения статистики осуществлено уже немало, и их результаты позволяют признать их удачными и полезными. Только не следует забывать, что статистика помогает науке о языке лишь в том случае, когда разумно сочетается с обычными, качественными методами анализа фактов и результатами такого качественного анализа. Сама возможность применения статистики предполагает умение выделять в речи и отождествлять одни и те же языковые элементы и категории, т. е. умение опознавать качества языка.

Можно, таким образом, сказать, что количественный подход к пониманию и определению стилей речи не снимает и не заменяет качественного и ведет, в перспективе, к обоснованию качественного. Но если все же количественный подход вычленить, то в аспекте его требований и возможностей стили речи могут получить следующее определение: стили речи — это разновидности ее структуры, соответствующие устойчивым особенностям внеязыковых и языковых условий ее построения и отличающиеся друг от друга существенными различиями частот языковых единиц и категорий (и различиями наборов таких частот), достаточными для их суммарного качественного опознавания на интуитивном уровне восприятия речи.

А это означает, что есть основания различать дифференциальные стилевые частоты и нейтральные стилевые частоты. Читатель уже знаком с выборочными частотами, полученными Н. Малиновской из произведений К. Симонова и М. Шолохова. Еще раз воспользуемся результатами этого опыта. Возьмем средние частоты некоторых морфологических и синтаксических явлений, сведем их в таблицу и знаками плюс и минус обозначим существенность или случайность расхождения частот[1].

Из 23 грамматических признаков речевых структур Симонова и Шолохова (напоминаю читателям, что из произведений того и другого писателя было взято по шести выборок длиной в 500 знаменательных слов каждая) в 14 статистика показала дифференцирующие стилевые частоты и в девяти — нейтральные.

Конечно, и эта таблица, и результаты положенного в ее основу опыта имеют скорее иллюстративное назначение, т. е. не должны рассматриваться как окончательные лингво-статистические характеристики и решения. Для того чтобы такие характеристики и решения получить, нужны новые и новые опыты, с охватом ими больших по объему текстовых массивов. Но и в качестве иллюстративного материала показанные читателю частоты и таблицы небезынтересны. Они убедительно говорят о больших возможностях, открываемых перед лингвистами (да и литературоведами) статистическим подходом к речевым стилям и соответствующей методикой.

[1] Необходимая статистическая обработка выборочных частот здесь, по условиям места, не показывается; не даются и исходные таблицы с выборочными частотами.



| Классы грамматических явлений | Авторы | | Существенность или случайность расхождения частот |
|---|---|---|---|
| | *Симонов* | *Шолохов* | |
| Имя существительное | 170 | 216 | + |
| Имя прилагательное | 49 | 77 | + |
| Местоимение | 73 | 39 | + |
| Имя числительное | 15 | 7 | + |
| Глагол | 111 | 77 | + |
| Причастие | 14 | 32 | + |
| Деепричастие | 4 | 12 | + |
| Наречие | 61 | 42 | — |
| Предлог | 86 | 87 | — |
| Союз | 76 | 41 | + |
| Подлежащее | 63 | 58 | — |
| Сказуемое | 104 | 73 | + |
| Связка | 10 | 4 | + |
| Обособленные слова | 19 | 24 | — |
| Однородные | 63 | 80 | — |
| Зависящие вправо от господствующего | 162 | 171 | — |
| Зависящие влево от господствующего | 153 | 190 | + |
| Вводные | 3 | 2 | — |
| Сложные предложения | 27 | 14 | + |
| Простые самостоятельные предложения | 12 | 20 | — |
| Сочинительные связи | 13 | 6 | + |
| Подчинительные связи | 25 | 7 | + |
| Бессоюзные связи | 10 | 12 | — |

В реальных речевых структурах дифференцирующие стили и нейтральные к стилям частоты сплетены и действуют на сознание читателя совместно. И хотя речевой стиль в целом интуитивно улавливается, замечается читателем или слушателем, составные элементы стиля, его сла-

гаемые интуицией обычно не вычленяются. Именно поэтому так важна статистика в изучении речевых стилей. Для описания стилей могут использоваться не только дифференцирующие частоты, но и существенно различные соотношения частот. Предположим, текст *А* дал на выборку в 500 знаменательных слов в среднем 200 имен существительных и 100 глаголов; а текст *Б* дал на такую же выборку в среднем 150 имен существительных и 95 глаголов. Частоты глаголов различаются несущественно, но этого нельзя сказать о расхождениях отношений имени существительного к глаголу (200 : 100 и 150 : 95). Каждый грамматический или лексический разряд слов имеет свою типовую, разрядовую семантику, которую можно уподобить окраске, определенному цвету. И вполне понятно, что общее впечатление, вызываемое в сознании читателя или слушателя воздействием речевого стиля, зависит не только от количества краски (занятого ею пространства), но и от тех соотношений цветов, которые использует писатель-художник, от свойственного ему цветового целого.

Одной из количественных характеристик стиля речи может служить и колеблемость частот. Если наблюдатель берет из текста ряд однородных и одинаковых по длине выборок и извлекает из этих выборок частоты наблюдаемых языковых явлений, то он обнаруживает большую или меньшую колеблемость частотного ряда. Величину колеблемости можно измерить либо при помощи критерия «хи-квадрат», либо при помощи коэффициента вариации. Но как бы эта колеблемость ни измерялась, она не может не интересовать лингвиста как одна из объективных характеристик описываемого речевого стиля и как своеобразно закодированная информация о совокупности стилеобразующих условий текста.

Определяя и оценивая методами и приемами статистики выборочные частоты, выборочные средние частоты, границы «действительных средних», колеблемость частотных рядов, лингвист получает объективную информацию, позволяющую намечать границы между функционально-тематическими, жанровыми, индивидуальными речевыми стилями, стилями литературных школ и направлений и иными.

Правда, если для решения многих прикладных задач описание речевого стиля может быть лишь статистическим, то для решения теоретических задач языкознания, а отчасти и прикладных статистическое описание — основание для перехода к более глубокому, качественному описанию стиля, для объяснения установленных при помощи статистики закономерностей структурными (внутриязыковыми) и внеязыковыми условиями функционирования языка и формирования структуры речевых последовательностей.



Язык, как сверхсложная структура, как большая система, не работает и не развивается сам в себе и сам для себя: его работа и его развитие есть результат постоянного взаимодействия, по крайней мере, трех систем: системы самого языка, системы сознания и системы объективной действительности. Можно извлекать из этого сложного триединства и язык, и сознание и рассматривать их как самостоятельные и самодовлеющие структуры. Но такое рассмотрение всегда оказывается ограниченным и рано или поздно приводит к необходимости вспомнить о том, что давным давно высказано К. Марксом: ни мысли, ни язык не образуют сами по себе особого царства, они только проявления действительной жизни. И в этом сложном своем аспекте язык также нуждается в исследовании, и статистика этому может помочь.

## ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ

1. Может быть, полезно задуматься о том, что дает уже в настоящее время статистическая методика лингвисту, какие проблемы позволяет решать более строго и объективно, чем это удавалось ранее. При этом хорошо бы увидеть и трудности, вызываемые своеобразием методики, и перспективы ее более широкого применения. Автор этой работы не вправе, разумеется, претендовать на освещение всего круга вопросов, которые возникают в области проблем и перспектив, и хотел бы высказать некоторые соображения о вещах, ему более близких.

Осуществленные и осуществляемые опыты применения статистики к решению лингвистических задач позволяют предполагать появление в недалеком будущем грамматик и словарей, описывающих не язык вообще, а его функциональные стили, т. е. типы его функционирования. Для того чтобы такие грамматики и словари появились, необходимо широкое изучение стилевой дифференциации языка

на всех его структурных уровнях, и прежде всего на уровнях лексики, морфологии и синтаксиса.

В современных грамматиках и словарях описываются слова и грамматические категории так, как будто нет никаких различий в структуре и функционировании языковых явлений в разных языковых стилях. Но теперь уже очевидно, что это не так. Глагол русского языка не один и тот же и по-разному работает в науке и публицистике, художественной литературе и деловой деятельности, на производстве и в быту людей. Неодинаковы структуры простого и сложного предложения и в особенности их функционирование в названных областях жизни.

Так почему бы не представить такие грамматики, в которых описание различных глагольных форм и значений будет включать и сведения о вероятностях этих форм и значений в разных стилях языка? А в ряду таких вероятностей были бы и нулевые, говорящие читателю о том, что такое-то значение или такая-то форма в таком-то стиле вообще не применяется. Подобным же образом описывались бы все грамматические явления нашего языка, включая и предложения различных структур.

Нетрудно себе представить преимущества такого описания грамматики перед ныне принятыми описаниями. Прежде всего, стало бы ясно, что многие грамматические явления, которым в настоящее время уделяется большое внимание в учебных курсах, практически не актуальны, потому что или вообще почти неактивны, или же активны в том или ином отдельном языковом стиле. Появилась бы, таким образом, реальная возможность отобрать из всего многообразия грамматических средств языка, изучаемых по традиции единым потоком, такие средства, которые реально функционируют в современной литературной речи.

Затем возникла бы возможность увидеть, как меняются варианты структуры отдельных грамматических категорий в случаях переключения этих категорий с обслуживания одной области деятельности общества на обслуживание другой области. Тем самым было бы достигнуто более полное, глубокое и правильное познание и узлов грамматического строя, и их функционирования, и языковых стилей.

Описание грамматического строя, включающее систематические вероятностные оценки, создавало бы неизвестное ныне представление о «работающей», функционирующей грамматике и тем самым о ее стилевых вариантах.



Подобное описание могло бы быть распространено и на лексику. Начало этому можно видеть в уже создаваемых частотных словарях, ориентированных на отдельные области науки и техники. Такие словари создаются, например, группой лингвистов, руководимой Р. Г. Piотровским. Эти словари очень нужны и для улучшения подготовки студентов по иностранным языкам, и для налаживания службы научно-технической информации с участием электронно-счетных машин. Конечно, в принципе нет непреодолимых помех и для создания частотных словарей, ориентированных на основные функциональные стили языка, как нет принципиальных помех и для дифференцированного по стилям описания лексики традиционными в лексикографии приемами. И можно, таким образом, представить время, когда наряду с общеязыковыми толковыми словарями будут существовать и служить людям толково-частотные словари художественной литературы, науки, публицистики, делопроизводства, быта и т. д.

Одним словом, в перспективе — создание, при активном и обязательном участии статистической методики, дифференцированного по стилям описания лексики, морфологии, синтаксиса, словообразования, да и фонетики, хотя фонемное устройство языка, по-видимому, менее чем другие структурные его области, дифференцировано по стилям в процессе развития и функционирования.

2. Противники применения точных методов в изучении художественного произведения не раз вспоминали приписываемую Пушкину мысль о том, что не дело поверять гармонию алгеброй, а значит — легкое умозаключение — не дело высокое искусство изучать при помощи низкой статистики.

Может быть, истины ради вспомнить, что же сказано у Пушкина? Сальери говорит:

. . . . . Труден первый шаг
И скучен первый путь. Преодолел
Я ранние невзгоды. Ремесло
Поставил я подножием искусству;
Я сделался ремесленник: перстам
Придал послушную, сухую беглость
И верность уху. Звуки умертвив,

Музыку я разъял, как труп. Поверил
Я алгеброй гармонию. Тогда
Уже дерзнул, в науке искушенный,
Предаться неге творческой мечты.
Я стал творить...

Так о чем же здесь сказано? О том, что *подножием искусству* для каждого хорошего музыканта (а Сальери хороший музыкант) оказывается ремесло, выработка навыков, познание, наука. Только создав подножие искусству, музыкант начинает творить. Поверять алгеброй гармонию необходимо для того, чтобы овладеть искусством — вот мысль Пушкина.

Но, конечно, все это ради истины в споре.

По существу же, дело, конечно, не в том, в каком именно смысле применены Пушкиным слова «поверять гармонию алгеброй» и признавал или не признавал правомерной такую поверку сам Пушкин. Дело совсем в другом: может или не может статистическая методика показать хотя бы некоторые особенности художественной речи и тем самым помочь ее более полному и глубокому пониманию? Если может, ее надо применять. Если не может — надо отвергнуть как непригодную для изучения специфики того языка, которым обслуживается художественная литература. Все дело в этом, и только в этом. Никто не объявлял количественную, статистическую методику универсальным орудием исследования всех сторон художественного творчества. Универсальных методических инструментов ни одна наука не знает. Тем более — никто не отвергал участия интуиции в исследовании художественного творчества. Речь идет лишь о требовании самой развивающейся науки — искать более строгие, доказательные, точные приемы исследования, включать в уже освоенный арсенал исследовательских инструментов новые, в особенности такие, которые обещают объективную проверку хотя бы некоторых гипотез, полученных на основе одной интуиции.

Дает ли, в этом смысле, что-нибудь статистическая методика лингвисту и литературоведу, изучающим художественную речь?

Взгляните, пожалуйста, на те немногие сведения о частотах и вероятностях в разных стилях языка и речи, которые были даны на страницах этой книги. Разве сами по себе эти сведения не говорят о том, что они вполне удовлетворительно и совершенно объективно информируют ученого-филолога о своеобразии художественной речи вообще и своеобразии художественной речи отдельных писателей в частности? Почему же нужно отвергать эту информацию? К тому же вполне возможно, что за сухими расхождениями и схождениями частот и вероятностей филолог со временем научится различать условия, их вызывающие, т. е. перейдет от структуры речи к структуре художественного содержания.

Почему же эту информацию нередко отвергают?

Трудно ответить на такой вопрос, познакомившись с возражениями некоторых литературоведов, упорно настаивающих на своем убеждении в непригодности статистики для изучения художественного произведения.

Но послушаем возражающих.

«Конечно, литература неразрывно связана с языком, а язык во многих отношениях поддается статистической обработке, тем не менее он входит в такое сложное взаимодействие с другими сторонами художественного творчества, которое представляет собой явление не статистического характера, и, стало быть, само математическое его изучение в отрыве от художественного процесса в целом не может привести к декларируемой сторонниками математической поэтики качественной оценке» [1].

«В целом ряде работ, анализирующих художественную речь, большое место занимают, например, отвлекающиеся от всей остальной реальности речи рассуждения о пропорции существительных, прилагательных и глаголов в тексте, о количестве метафор и сравнений, о соотношении подчинительных и сочинительных конструкций, о характере аллитераций и т. п. Все эти моменты могут, конечно, играть существенную роль, но, будучи вырваны из целостности речи, из всесторонней взаимосвязи стилевого единства, они очень мало что объясняют. А целостность эта в значительной степени основывается на своеобразном ритме речи, характерном для произведения. Он, бесспорно, накладывает глубокую печать на общую структуру фразы, на ее звуковой строй и опосредованно на лексико-фразеологическую материю речи...

[1] Л. И. Тимофеев. Число и чувство меры в изучении поэтики. В кн.: «Слово и образ». М., 1964, стр. 287.

Каждая форма художественной речи должна быть понята не как совокупность тех или иных средств, но как целостная система вполне определенных и уже неподменимых средств; удаление какого-либо элемента (а тем более связи — ритма, основных принципов строения фразы) меняет и разрушает все» [1].

«М. В. Карпенко подсчитала, сколько предложений с разным количеством слов содержится в первой части «Войны и мира». Эти подсчеты в общем подтверждают то положение (впрочем, бесспорное), что у Толстого весьма много сложных синтаксических образований. Однако таблицы такого рода непременно требуют сопоставлений. Но и сопоставления в своей ценности были бы весьма ограничены. Что из того, что у Карамзина синтаксис мог бы оказаться в количественном отношении близким синтаксису Толстого? Ведь стилистическое назначение сложных синтаксических форм у этих двух писателей совсем разное, да и разные, конечно, эти формы, не по количеству слов в предложении, а по своему грамматическому строю» [2].

Что же нам говорят противники меры и числа в изучении художественной речи?

Что язык в художественном произведении входит в сложное взаимодействие с другими сторонами художественного творчества, это сложное взаимодействие не подчинено законам статистики, и изучать его в отрыве от художественного процесса в целом нельзя.

Что соотношения частот различных элементов языка могут играть существенную роль, но их нельзя вырывать из целостной речи, а целостность речи основывается, оказывается, на ее своеобразном ритме.

Что художественная речь может быть понята только как целостная система, иначе разрушается все.

Что сами по себе количественные данные немного стоят, все дело в функциях языковых единиц.

Но позвольте, разве кто-нибудь уже доказал, что художественное целое во всей сложности взаимодействий его сторон и элементов не подчиняется вероятностным законам? А если подчиняется? Пусть не во всей своей структуре, пусть в ее каких-то сторонах и узлах,— разве этого недостаточно, чтобы сделать необходимым применение статистики?

На каком реальном основании вновь и вновь воздвигается пугало отрыва изучаемых речевых структур от художественного целого? Разве кто-нибудь уже сумел охватить исследовательским оком это целое? И разве уже существует методика целостного анализа художественного произведения? А без такой методики не подменяется ли нередко ожидаемый анализ словесными декларациями о нем? Да и может ли наука успешно познавать сложный объект, не вычленяя его отдельных сторон и не отвлекаясь по необходимости от всей совокупности реальных связей объекта?

Кем, когда, на каком материале установлено, что целостность речи основывается на своеобразном ритме? Ведь это слишком ответственное и слишком бездоказательное заявление, чтобы оно могло быть включено в число аргументов против статистики. Как раз статистика уже показывает, что стилевое своеобразие речевых структур создается участием в них языковых единиц различных уровней, а вовсе не одной ритмомелодики.

Дело не в количестве, а в функциях... А если у количества есть свои функции? Активны прилагательные — одни функции, пассивны — другие? Да ведь и наши суждения о функциях грамматических и иных средств языка не имеют подчас доказательной силы и нуждаются в проверке...

После всех этих недоуменных вопросов обратимся еще раз к свидетельству фактов. В одном из опытов [1] были получены частоты некоторых грамматических явлений в авторской художественной речи А. С. Пушкина и М. Ю. Лермонтова. Общая длина выборок из текстов каждого писателя равняется 10000 знаменательных слов, а длина одной выборки — 500 слов. Обработка исходных статистических данных, полученных в опыте, позволяет получить ряд статистических обобщений, характеризующих речевое и художественное творчество двух основоположников реалистической русской прозы.

В среднем на одну выборку, т. е. на 500 знаменательных слов текста, в прозе Пушкина приходится 182 имени существительных, 46 имен прилагательных, 100 место-

[1] В. В. Кожинов. Слово как форма образа. В кн.: «Слово и образ». М., 1964, стр. 16.

[2] А. В. Чичерин. Заметки о стилистической роли грамматических форм. В кн.: «Слово и образ». М., 1964, стр. 100.

[1] Данные этих опытов отражены в дипломных работах студентов ГГУ Г. Сальгиной и С. Оболяевой.

имений, 11 глаголов, 28 наречий; в прозе Лермонтова — 164 имени существительных, 62 имени прилагательных, 109 местоимений, 101 глагол, 43 наречия.

На основе этих сведений о средних частотах частей речи можно получить сведения о соотношениях частей речи в текстах Пушкина и Лермонтова. Так, разделив количество прилагательных на количество существительных, мы получим десятичную дробь, которая и покажет, в каком количественном соотношении находятся прилагательные и существительные. Подобным же образом можно получить сведения о количественных соотношениях глаголов и существительных, местоимений и глаголов, глаголов и наречий и т. д. Такие соотношения нередко нагляднее и глубже показывают специфику речевых структур, чем это делают разрозненно взятые частоты. И это можно понять. Ведь соотношения частей речи или других элементов языка говорят о степени густоты качественной окраски, даваемой прилагательными существительным, процессуальной окраски, даваемой глаголами существительным и местоимениям, качественной окраски, даваемой наречиями глаголам, и т. д.

Каковы же соотношения некоторых частей речи в текстах Пушкина и Лермонтова?

Вот небольшая таблица.

| Соотношение | *Пушкин* | *Лермонтов* |
|---|---|---|
| Прилаг. : существ. | 0,25 | 0,38 |
| Местоим. : существ. | 0,55 | 0,66 |
| Местоим. : глагол | 0,91 | 1,07 |
| Существ. : глагол | 1,65 | 1,63 |
| Наречие : глагол | 0,26 | 0,42 |
| Наречие : прилаг. | 0,63 | 0,70 |
| Прилаг. : глагол | 0,42 | 0,61 |
| Глагол : существ. и прилаг. | 0,39 | 0,37 |
| Наречие и прилаг. : существ. и глагол | 0,24 | 0,40 |

Несмотря на то что показанные в таблице величины неокончательные, приближенные, колеблющиеся в реальном тексте [1], они говорят о многом.

[1] Пределы колебания, чтобы не осложнять изложение, не показываются.



Если применить одну из описанных ранее процедур сравнения долей (или частот), можно установить, что соотношения «прилагательное : существительное», «местоимение : существительное», «местоимение : глагол», «наречие : глагол», «прилагательное : глагол», «наречие и прилагательное : существительное и глагол» существенно различны в речи Пушкина и Лермонтова; соотношения «существительное : глагол», «наречие : прилагательное», «глагол : существительное и прилагательное» различаются в текстах Пушкина и Лермонтова несущественно. И не являются ли расхождения показанных соотношений частей речи выражением закономерных различий хотя бы некоторых сторон творческого видения и изображения действительности писателями?

Таблица наглядно убеждает читателя в том, что Лермонтов видит и изображает мир, действительность в большем разнообразии качественных характеристик, красок, признаков, чем это делает Пушкин: в среднем у Лермонтова 39 имен существительных из каждых 100 получают признаки прилагательных, а у Пушкина таких существительных всего 25; значит, речь и мышление Пушкина предметнее, Лермонтова — «качественнее». О сходной особенности речи и мышления двух авторов говорит и соотношение «наречие : глагол»: в среднем у Лермонтова на 100 глаголов приходится 42 наречия, а у Пушкина — всего 26; а это означает, что речь и мышление Лермонтова активнее окрашивают воспринимаемые процессы, чем это делают речь и мышление Пушкина. Интересно и то, что соотношение «существительное : глагол» и «глагол : существительное и прилагательное» оказываются у Пушкина и Лермонтова статистически равными, что позволяет опровергать гипотезу об особой глагольности речевого стиля Пушкина.

Посмотрим теперь некоторые данные из области синтаксиса. В прозе Пушкина, в авторской речи, на каждые 500 знаменательных слов приходится в среднем около 26 простых самостоятельных предложений, около 20 сложных, около 53 простых в составе сложных, около 10 сочинительных связей, около 14 подчинительных и около 8 бессоюзных связей между предложениями в составе сложных. В прозе Лермонтова простых самостоятельных предложений около 11 на 500 знаменательных слов, сложных — око-

ло 22, простых в составе сложных — около 70, сочинительных связей — около 10, подчинительных — около 18, бессоюзных — около 21. Построим на основе этих данных небольшую таблицу соотношений.

| Соотношение | *Пушкин* | *Лермонтов* |
|---|---|---|
| Простые предложения : сложные предложения | 1,30 | 0,50 |
| Простые самостоятельные : простые в сложных | 0,49 | 0,16 |
| Сочинительные связи : подчинительные связи | 0,72 | 0,56 |
| Подчинительные связи : бессоюзные связи | 1,76 | 0,86 |
| Бессоюзные связи : подчинительные и сочинительные связи | 0,33 | 0,75 |

Что называется, невооруженным глазом видно существенное различие показанных соотношений в речи Пушкина и в речи Лермонтова. И не говорят ли эти различия о том, что речь и мышление Лермонтова слитнее, синтетичнее (сложных предложений вдвое больше, чем простых), чем речь и мышление Пушкина (сложных предложений меньше, чем простых); логические отношения между единицами мысли и речи в прозе Лермонтова менее определенно выражены, чем в мысли и речи Пушкина, об этом убедительно говорит соотношение союзных и бессоюзных связей; речь Лермонтова сложнее, структурно богаче и разнообразнее пушкинской прежде всего в области грамматических связей между предложениями, в области структуры предложений, ритмомелодики и интонации, которые тесно связаны со структурой предложения, и т. д. Если вспомнить и морфологические соотношения, можно сказать, что речь Лермонтова субъективнее речи Пушкина: Лермонтов активнее выражает свою оценку фактов действительности прилагательными, наречиями, бессоюзными связями предложений, оставляющими большие возможности варьирования в понимании, восприятии отношений между явлениями жизни.



Закономерные отличия прозаической речи Лермонтова от речи Пушкина, хотя эти отличия и были взяты лишь на абстрактных уровнях морфологии и синтаксиса, могут осознаваться как следствия особенностей лермонтовского реализма. Учась у Пушкина, развивая его речевые и художественные принципы и традиции, Лермонтов-прозаик вырабатывает собственные принципы в области прозаического речевого творчества, закладывает основания новых традиций, отличающихся от пушкинских. В 40-е годы XIX в. в русской реалистической литературе возникают, складываются д в е т е н д е н ц и и, д в е т р а д и ц и и в р а з в и т и и п р о з а и ч е с к о й р е ч и — п у ш к и н с к а я и л е р м о н т о в с к а я. И русские прозаики XIX в. (а возможно, и XX) либо развивают и обогащают пушкинскую или лермонтовскую речевую тенденцию, либо совмещают их, либо, наконец, их преодолевают, создавая новые тенденции, захватывающие отдельные участки речевой структуры, а затем и структуру отдельных стилей в целом, обслуживающих нужды художественной литературы. И развитие, и преобразование тенденций можно установить при помощи статистической методики.

Таким образом, даже беглый взгляд на применение статистики в изучении художественной речи и художественного творчества обнаруживает большие возможности, находит новые исследовательские проблемы и пути движения науки, изучающей художественную литературу. Чем активнее и полнее будет испытываться статистическая методика в роли союзницы и помощницы других методик филологами, тем яснее будут становиться перспективы и границы ее применения.

3. Можно надеяться и на то, что статистика будет хорошей союзницей филологии в тех случаях, когда оказывается необходимым решать вопросы речевой культуры в широком смысле, т. е. имея в виду не только обиходную речь, но и речь художественную, научную, публицистическую, деловую и иные ее типы и виды.

Допустимо предположение о том, что структура речи и моделируемая речью структура мышления обладают некоторыми свойствами или качествами, которые могут быть названы формальными, необходимыми при выявлении любого содержания, и которые могут быть выражены в тех или иных числовых показателях.

Так, если признать, что каждое простое предложение

передает в конкретной речевой цепи одно суждение, то можно будет ввести понятие и термин «темп речи-мысли» и измерять его отношением числа знаменательных слов к числу простых предложений; ведь чем меньше слов приходится на одно предложение, тем чаще предложения (а значит, и мысли) сменяют друг друга, т. е. тем больше темп речи. Однако для его измерения лучше пользоваться не просто соотношением числа слов и числа предложений, а некоторым преобразованием и обогащением этого соотношения: примем в качестве гипотезы, что темп речи, принимаемый за единицу, мы получаем тогда, когда на одно предложение приходится семь знаменательных слов (величина несколько условная, но все же учитывающая и речевую реальность); примем и другую гипотезу — об увеличении темпа речи-мысли в зависимости от числа главных членов, вошедших в двусоставное предложение: ведь чем больше главных членов, тем больше предикативных связей, а значит, и суждений. Приняв во внимание все сказанное, получаем формулу: $T = \frac{7N(P+S)}{2nN_2}$, где $N$ — число простых предложений, $N_2$ — число двусоставных, $n$ — число знаменательных слов, $P$ — число подлежащих, $S$ — число сказуемых, 7 и 2 — коэффициенты, обеспечивающие одну единицу темпа при семи словах в одном предложении и при одном подлежащем и одном сказуемом в каждом предложении.

Возьмем два конкретных примера.

Из русской сказки:

*«Захотелось волку посмотреть на Котофея Ивановича, да сквозь листья не видать, и начал он прокапывать над глазами листья. Кот услыхал, что лист шевелится, подумал, что это мышь, да как кинется — и вцепился когтями в волчью морду. Волк вскочил — да бежать! А кот сам испугался и бросился прямо на дерево, где медведь сидел.*

*«Ну, — думает медведь, — увидал меня!»*

*Слезать-то некогда, вот он и кинулся с дерева наземь — все печенки отбил, вскочил — да бежать. А лисица вслед кричит:*

*— Вот он задаст вам! Погодите!*

*С той поры все звери стали кота бояться. А кот с лисой запаслись на целую зиму мясом и стали жить да поживать, и теперь живут, хлеб жуют».*

Из научной статьи:

*«Творческие возможности социалистического реализма*



*так же, как и его достижения, заключены, разумеется, не только в масштабности охвата действительности, но и в глубине и разносторонности раскрытия жизни, в многообразии его национальных, индивидуальных проявлений, в богатстве художественных форм. Реализация этих возможностей, естественно, зависит прежде всего от таланта художников, характера связей их с социальной действительностью, но не в малой степени она зависит и от исторического осмысления современного литературного процесса, от преодоления ложных идей и предрассудков, которые существуют до сих пор в литературной теории».*

Применив нашу формулу, получим: темп речи-мысли в текстовом отрывке из сказки равен 2,39, а в отрывке из научной статьи — всего 0,42. Это значит, что речь в сказке развертывается быстрее почти в 6 раз (5,7), чем в научной статье. И это понятно: в сказке мысли и выражающие их высказывания просты по своей структуре, поэтому они быстрее, легче выстраиваются в динамическую последовательность; в научной статье структура мысли-речи намного сложнее, а поэтому каналы сознания пропускают единицы такой речи-мысли затрудненнее и медленнее.

Большая или меньшая активность в речи предлогов и союзов может быть положена в основу понятия и термина «связанность речи-мысли»: чем больше союзов и предлогов приходится на одно самостоятельное, простое или сложное предложение, тем больше связанность речи. Взяв отношение числа предлогов и союзов к числу самостоятельных предложений, мы получим некоторый показатель, коэффициент связанности. Условимся видеть этот коэффициент равным единице тогда, когда на одно предложение приходится три связывающих элемента, т. е. предлога и союза. При таком условии формула связанности получит следующий вид:

$Св. = \frac{П + С}{3N}$, где $П$ — число предлогов, $С$ — число союзов, $N$ — число самостоятельных, «от точки до точки», предложений.

Наши два опытных текста имеют такие коэффициенты связанности: текст сказки — 0,77, текст научной статьи — 3,0, т. е. связанность во втором тексте в 3,9 сильнее, чем в первом.

В структуре предложения (а точнее — высказывания) семантика подлежащего и сказуемого уточняется всеми

второстепенными членами, вошедшими в прямые или косвенные грамматические связи с одним из главных. Отсюда возможность принять понятие-термин «уточненность речи-мысли». Величина уточненности может измеряться отношением числа второстепенных членов к числу главных: $У = \frac{n_{вт}}{n_{гл}}$.

Отрывок из текста сказки дает коэффициент уточненности речи, равный 0,63, а отрывок из научного текста — коэффициент, равный 4,5! Уточненность речи сказки оказалась в нашем опыте в семь с лишним раз слабее, чем уточненность речи научной!

Обособленные, однородные и вводные члены предложения ослабляют его внутренние связи, как бы расчленяют структуру простого предложения своим вторжением. Поэтому можно принять понятие-термин «расчлененность единицы речи-мысли» и измерять эту расчлененность отношением числа обособленных, однородных и вводных членов к числу простых предложений: $Р = \frac{Об. + О + У}{N}$, где *Об.* — число обособленных, нераспространенных и распространенных членов; *О* — число однородных, нераспространенных и распространенных членов; *У* — число вводных, нераспространенных и распространенных членов; *N* — число простых самостоятельных предложений и простых предложений в составе сложных.

Отрывок из текста сказки дает коэффициент расчлененности, равный 0,9; отрывок из текста научной статьи — коэффициент, равный 4. Расчлененность единицы речи сказки оказалась в нашем опыте в 4,4 раза меньше, чем расчлененность единицы речи научной.

Члены предложения внутри высказываний могут быть связаны либо контактной, либо дистантной зависимостью:

Летят перелетные птицы
В осенней дали голубой.

Слова *перелетные* и *птицы*, *осенней* и *дали*, *дали* и *голубой* связаны контактно, слова *летят* и *птицы*, *летят* и *в дали* — дистантно. Отношение числа дистантных связей к числу контактных назовем «разрывностью речи-мысли»,



и это понятие-термин свяжем с соответствующей формулой: $Рз. = \frac{n_д}{n_к}$, где $n_д$ — число дистантных связей, $n_к$ — число контактных связей между словами.

Показатель разрывности в опытном куске текста из сказки равен 0,7, а в куске научного текста — 1,3, т. е. почти в два раза больше.

По предшествующему изложению читатель уже знаком с однородностью и неоднородностью текста речи. Допустимо говорить и об «ровности речи-мысли», связывая этот термин и это понятие с теми величинами, которые показывает критерий «хи-квадрат» или коэффициент вариации. Чем больше «хи-квадрат» (или коэффициент вариации), тем меньше, слабее ровность речи-мысли. Очевидно, что соответствующие характеристики речи разных стилей и разных авторов и жанров могут существенно обогатить наши знания о формальных свойствах и качествах обширных речевых структур. Очень возможно, что необходимые для этого исследования внесут много неожиданного в описание речи, в традиционные о ней теоретические представления.

Морфологические классы слов (части речи) дают основание ввести в обиход понятия-термины «предметность речи-мысли», «качественность речи-мысли», «действенность (динамизм) речи-мысли».

Предметность может измеряться отношением числа имен существительных и местоимений-существительных к числу прилагательных и глаголов: $П = \frac{С + Мс}{П + Г}$. Качественность можно измерить отношением числа прилагательных и наречий к числу имен существительных и глаголов: $К = \frac{П + Н}{С + Г}$.

Действенность (динамизм) допустимо определять через отношение глаголов к именам и местоимениям: $Д = \frac{Г + Пч + Дч}{С + П + М}$. В число глаголов можно включать или не включать причастия и деепричастия; в число имен придется, очевидно, включить имена существительные, имена прилагательные, возможно, и наречия. Варианты решения должны быть точно приняты в процессе эксперимента.

Вот небольшая таблица, показывающая предметность,

качественность и динамизм речи-мысли в двух экспери ментальных отрывках текста:

| Формальное качество речи | Текст сказки | Текст научной статьи |
|---|---|---|
| Предметность | 0,83 | 2,07 |
| Качественность | 0,07 | 0,36 |
| Динамизм | 0,97 | 0,12 |

Хорошо заметны резкие различия количественных показателей предметности, качественности и динамизма речи-мысли в сказке и научной статье. Может быть, небезынтересно узнать, что предметность речевого отрывка из научной статьи в 2,57 раза больше, чем предметность речевого отрывка из сказки; качественность, характеризующая конкретизацию прилагательными и наречиями существительных и глаголов, больше в научной статье в 5,14, чем в сказке; динамизм речи в научной статье оказался в нашем опыте в 8,08 раза меньше, чем в сказке.

Конечно, вошедшие в таблицу показатели предметности, качественности и динамизма речи-мысли носят лишь предварительное и иллюстративное назначение: чтобы стать достоверными и надежными, показатели должны опираться на ряды выборок и обрабатываться в соответствии с требованиями математической статистики. Однако и в своем предварительном назначении показатели таблицы очень поучительны: они уверенно говорят о серьезных, существенных, закономерных различиях в интенсивности оцениваемых качеств речи в разных ее видах и типах. Отсюда следует целесообразность их изучения под описываемым углом зрения.

По-видимому, была бы оправдана попытка найти формулы, позволяющие оценивать сложность и богатство (разнообразие) речи-мысли.

Допустимо думать, что сложность речи увеличивается с увеличением числа знаменательных слов, приходящихся на одно простое предложение, и с увеличением числа простых предложений, приходящихся в среднем на одно самостоятельное предложение («от точки до точки»). Примем за некий эталон синтаксической простоты речь, в которой на одно простое предложение приходится два знаменательных слова и каждое самостоятельное предложение — простое: «Дождь перестал. Небо очистилось. Ветер утих. Повеяло теплом. Запели птицы».



Несложные соображения позволят предложить следующую формулу для вычисления коэффициента синтаксической сложности речи: $Сл. = \frac{n \cdot к}{2к \cdot N} = \frac{n}{2N}$, где $n$ — число знаменательных слов, $к$ — число простых предложений, $N$ — число самостоятельных предложений, простых или сложных.

В одном из опытов были взяты куски текстов из произведения Л. Леонова «Русский лес» и из рассказа К. Паустовского «Дождливый рассвет»; совокупная выборка из каждого произведения равнялась 1000 знаменательных слов. В романе Л. Леонова на 1000 слов оказалось 105 простых предложений и 39 самостоятельных; в рассказе К. Паустовского — 178 простых и 136 самостоятельных.

Применение формулы синтаксической сложности речи дало для текста Л. Леонова коэффициент, равный 12,7, для текста К. Паустовского — 3,64; это значит, что речь Паустовского в 3,5 раза проще речи Леонова. Повторный опыт строился на выборках в 600 знаменательных слов из романа «Русский лес» и из рассказа «Корзина с еловыми шишками». Получены коэффициенты: для текста Л. Леонова — 9,75, для текста К. Паустовского — 4,15, т. е. в 2,34 раза меньше.

Можно формулу для вычисления коэффициента сложности речи-мысли несколько видоизменить, опираясь на предположение о том, что наличие в структуре простого предложения обособлений увеличивает сложность; допустим при этом, что одно обособление по своей усложняющей силе равно половине простого предложения в составе сложного. Формула примет вид: $Кф. = \frac{n \cdot (2к + 0)}{4кN}$, где 0 — число обособлений.

В этом варианте формула синтаксической сложности дает следующие коэффициенты для показанных ранее читателю кусков текста сказки и научной статьи: сказка — 3,9, научная статья — 15,5, т. е. почти в четыре раза больше!

Попытаемся измерить сложность ритмомелодической организации стихотворной речи, предполагая, что ее сложность зависит от соотношения безударных и ударных сло-

гов и от числа знаменательных слов в одной стихотворной строке; примем при этом за эталон сложности стихотворную речь вида:

Последний лед
Река несет.
Скворец поет:
Весна идет!

Получим формулу: $Сл. = \frac{T_б \cdot n}{T_у \cdot 2l}$, где $T_б$ — количество безударных слогов, $T_у$ — количество ударных слогов, $n$ — количество знаменательных слов, $l$ — количество строк.

Если принять во внимание то, что количество ударных слогов практически равняется количеству знаменательных слов, формулу можно упростить: $Сл. = \frac{T_б}{2l}$.

Для небольшого опыта были взяты куски стихотворного текста длиной в 500 знаменательных слов из од А. П. Сумарокова, М. В. Ломоносова, Г. Р. Державина и из поэмы «Полтава» А. С. Пушкина. Применение формулы дало следующие коэффициенты: Сумароков — 2,61; Ломоносов — 2,65; Державин — 2,45; Пушкин — 2,6. В тех же самых кусках текста коэффициенты синтаксической сложности заметно понижались от Сумарокова к Пушкину: Сумароков — 5,3; Ломоносов — 4,7; Державин — 4,2; Пушкин — 3,72.

Ничто не мешает вычислить коэффициенты, измеряющие совмещение сложности синтаксической и сложности ритмомелодической по формуле: $Сл. = \frac{n \cdot T_б}{4Nl}$. Объединенный коэффициент для текста Сумарокова оказался в опыте равным 13,8; для текста Ломоносова — 12,5; для текста Державина — 10,3; для текста Пушкина — 9,7. Картина если и не вполне пока надежная, но, во всяком случае, любопытная!

Интересно, что объединенный коэффициент сложности для 10 строк пушкинского стихотворения «Буря мглою»... равен 15,8, а для стихотворения «Я вас любил» — 29,11.

Кто знает, не дадут ли новые и новые опыты измерения сложности речевых структур такие результаты, которые заметно обогатят и изменят существующие представления о некоторых объективных свойствах речи различных типов? Конечно, при этом могут измениться и сами инструменты



для измерения сложности, здесь лишь намеченные в форме гипотезы.

Лексическое богатство мысли-речи может оцениваться отношением числа примененных лексем (отдельных слов) к числу их употреблений, т. е. к длине текста. Например, если в некотором тексте *А* длиной в 1000 словоупотреблений оказалось 100 разных слов (лексем), а в тексте *Б* такой же длины — 150 разных слов, мы можем сказать, что лексическое богатство второго текста больше.

Замечено, что существует, по-видимому, закономерная зависимость между активностью слов и тем местом, которое наиболее активные слова занимают в тексте, в речевом массиве. Если речевой массив достаточно велик, исчисляется сотнями словоупотреблений, то около 80% его совокупной длины занимаются (или покрываются) примерно двумя тысячами самых активных слов (лексем). Иначе говоря, если по достаточно большому речевому массиву мы составим частотный словарь и расположим в нем слова в порядке убывания их частот, т. е. их активности, то окажется, что первые 2000 слов нашего словаря занимают примерно 0,8 всего речевого массива, заполняют около 0,8 его длины.

Так, по данным Э. А. Штейнфельдт, около 2000 наиболее частых слов занимают такую долю в обследованных текстах:

| | |
|---|---|
| в рассказах и повестях советских писателей для среднего школьного возраста | —0,81 |
| в переводах сказок и повестей зарубежных писателей для того же возраста | —0,80 |
| в повестях, рассказах, поэмах для среднего и старшего возраста (советские писатели) | —0,78 |
| в пьесах советских авторов | —0,79 |
| в газетах «Пионерская правда» и «Комсомольская правда» | —0,78 |
| в журналах «Юность», «Смена», «Вокруг света» | —0,77 |
| в радиопередачах для молодежи | —0,77 |

Эти и другие интересные данные, содержащиеся в работе Э. А. Штейнфельдт [1], предназначены вовсе не для

[1] См.: Э. А. Штейнфельдт. Частотный словарь современного русского литературного языка. Таллин, 1963, стр. 73—88.

характеристики лексического богатства мысли-речи, однако они полезны и при осмыслении этого вопроса.

Ученые Ж. Эсту, Э. Кондон, Дж. Ципф, Б. Мандельброт и другие попытались уловить зависимость между порядковым номером слова в частотном словаре (так называемый «ранг» слова) и вероятностью этого же слова. Дж. Ципф и Б. Мандельброт предложили формулу, приближенно показывающую эту зависимость: $P_r = Kr^{-\gamma}$, где $P_r$ — вероятность (или относительная частота) слова, $r$ — ранг (порядковый номер) слова, $K$, $\gamma$ — постоянные величины, коэффициенты, устанавливаемые на основе опыта и теории. Дж. Ципф принял величину $K$ равной 0,1, а величину $\gamma$ равной 1. Проверка «закона Ципфа — Мандельброта» показала, что этот «закон» лишь очень приближенно соответствует языковой реальности и нередко ее заметно искажает. «Константы» $\kappa$ и $\gamma$ оказываются неконстантными и требуют варьирования в зависимости от стиля, жанра, эпохи и т. д. И все же «закон Ципфа — Мандельброта» интересен как попытка более или менее точно оценить существующую в текстах связь между частотой (а значит, и вероятностью) слова и его местом в ранжированном частотном словаре.

Однако и закон Ципфа не может быть прямо использован для характеристики лексического богатства речи, хотя он и улавливает какое-то отношение числа различных слов к числу их употреблений.

Возьмем прямое отношение числа разных слов ($L$) к числу их употреблений ($n$) в достаточно больших массивах речи. В текстах Пушкина $\frac{L}{n} = \frac{21\,197}{544\,777} = 0{,}039$; в текстах, на основе которых составлен словарь Штейнфельдт, $\frac{L}{n} = \frac{24\,224}{400\,000} = 0{,}065$; в текстах по органической химии $\frac{L}{n} = \frac{8000}{10\,000} = 0{,}08$; в текстах, на основе которых составляется частотный словарь преподавателями ЛГУ, $\frac{L}{n} = \frac{14\,208}{120\,474} = 0{,}117$ [1].

В идеале в речи не должно быть ни одного повторяющегося слова и $\frac{L}{n} = 1$. Значит, чем меньше десятичная дробь, измеряющая отношение $L$ к $n$, тем меньше лексическое богатство речи-мысли. Получается, что тексты Пушкина беднее лексически текстов, послуживших основой для словаря Штейнфельдт и даже текстов по органической химии!

Но, кажется, мы сравнили показатели речевого лексического богатства не очень корректно: ведь длина текстов Пушкина значительно превосходит длину текстов по органической химии (544777 и 10000 словоупотреблений), да и тексты, служившие основой словаря Штейнфельдт, короче пушкинских. Между тем хорошо известно, что увеличение числа разных слов, обслуживающих текст, не пропорционально увеличению длины текста, а испытывает отставание.

Поэтому и нельзя сравнивать показатели речевого лексического богатства так, как это было сделано только что. Нужна иная, более гибкая методика. Ею могло бы стать последовательное измерение отношения $\frac{L}{n}$ для некоторого отрезка текста, затем его удвоенной, утроенной, учетверенной и т. д. длины. Например, берем кусок текста длиной в 200 словоупотреблений и узнаем отношение $L$ к $n$, затем прибавляем еще кусок текста такой же длины и измеряем $\frac{L}{n}$ в уже удвоенном по длине отрезке текста и т. д. В результате мы получим данные не только об отношении $\frac{L}{n}$, но и о постепенном уменьшении этого отношения по мере удлинения текста. И то и другое может быть выражено целыми числами и дробями, а также кривыми линиями графиков, описывающих динамику насыщения речи словами.

Можно представить и иную, более простую и обычную для статистики методику. Из текстов двух языковых стилей (например, художественно-прозаическая речь и речь публицистическая) или стилей авторских экспериментатор берет по равному числу выборок равного объема (скажем, по 10 выборок в 1000 знаменательных слов каждая) и в каждой выборке подсчитывает число лексем; затем частоты обрабатываются описанным ранее способом и сравниваются; в результате — объективная статистическая картина различий в речевом лексическом богатстве двух, а затем и многих текстов.

[1] См.: Л. Н. Засорина. Автоматизация и статистика в лексикографии. Изд. Ленинградского университета, 1966, стр. 59.

Таким образом, статистика позволяет ставить опыты по измерению таких формальных качеств речи-мысли, которые были названы (может быть, и не очень удачно) темпом, связанностью, уточненностью, расчлененностью, разрывностью, ровностью, предметностью, качественностью, динамизмом, сложностью и богатством. Каждое из этих качеств понимается автором только в том узком смысле, который был сообщен читателям. Но и в этом узком смысле каждое из качеств речи-мысли заслуживает внимания лингвистов, литературоведов, психологов. Вероятно, изучение этих качеств внесет существенные поправки в высказанные автором гипотезы, а возможно, и отвергнет некоторые из них. Однако независимо от этого останется задача поиска новых подходов филолога к речевым структурам и новых методик их изучения.

4. Привлекает возможность применения статистики в изучении речи школьника, ее структуры и ее развития. Нужда в решении соответствующих задач очевидна. Достаточно сослаться хотя бы на то, что наша школа до сих пор не имеет сколько-нибудь надежных, систематических, объективных сведений о том, что такое хорошая, удовлетворительная и слабая речь школьника I, V, VIII, X классов, каков необходимый для успешных занятий в каждом классе запас слов, грамматических форм и конструкций, интонаций и т. д., как изменяется с возрастом и общим развитием школьника структура его речи, как влияет на развитие речи школьника урок, книга, радио, семья, улица. Вопросов таких много, они ждут ответа или хотя бы, на первых порах, пристального внимания. Не может ли пригодиться и в этой области статистическая методика?

Сошлюсь на один конкретный пример.

Летом 1967 г. успешно защитила дипломную работу выпускница Горьковского университета, заочница, учительница О. Г. Бугрова. Тема дипломной работы — «Опыт изучения определительных отношений в письменной речи учащихся пятых и восьмых классов». Работа была выполнена на основе эксперимента, сопровождавшегося статистической обработкой его результатов. Учащиеся пятых и восьмых классов писали сочинения повествовательного и описательного содержания. Сочинения, в зависимости от совокупности интуитивно учитываемых признаков, делились учителем на две группы — группу хороших и группу слабых. Тексты сочинений обрабатывались по стандартной программе, предусматривающей принадлежность слов к той или иной части речи и морфологическую природу согласованных и несогласованных определений, а также элементарную группировку прилагательных по семантическим типам. В результате эксперимента установлены некоторые существенные различия в структуре определительных отношений: а) между описанием и повествованием, б) между речью хорошей и речью слабой, в) между речью учеников V и VIII классов.

К сожалению, здесь нет возможности показать конкретные данные, полученные учительницей О. Г. Бугровой. Но я процитирую ее слова, говорящие о ее отношении к примененной методике: «Думается, что... вероятностно-статистическая методика позволила отчасти решить нашу задачу. Именно применение некоторых инструментов математической статистики дало возможность обнаружить некоторые тенденции в развитии определительных отношений: меньшую или большую потребность в использовании разных грамматических категорий, закономерность или случайность в их употреблении, равномерность или неравномерность развития.

Таким образом, вероятностно-статистическая методика поможет исследователям детской речи установить некоторые тенденции в развитии речи учащихся, что необходимо для решения практически действенной задачи обогащения речи школьника» (стр. 105 дипломного сочинения).

В этих высказываниях очень интересно признание учителем практической пользы той методики, которая была испытана, признание глубокой связи между изучением закономерностей развития речи учащихся и работой школы по воспитанию речевой культуры своих питомцев.

5. Применение статистики обещает много нового в области атрибуции литературных текстов.

Ведь если речевому творчеству определенного автора свойственны устойчивые, вскрываемые статистикой закономерности в использовании тех или иных единиц и категорий языка и если эти закономерности удалось установить, то, по-видимому, не покажется слишком сложной задача проверить, не написано ли именно этим автором некоторое безымянное сочинение. Если в этом сочинении определен-

ный ряд наблюдаемых фактов подчинен тем же статистическим закономерностям, имеет те же статистические характеристики, что и в речи предполагаемого автора, то почти наверное можно будет сказать, что авторство установлено.

Вспомните, пожалуйста, читатель, те ряды частот, которые измеряют активность частей речи в произведениях Симонова и Шолохова. Ведь если бы, предположим, нужно было установить, кем, Симоновым или Шолоховым,— написано некоторое безымянное произведение, достаточно было бы взять из него несколько проб текста, статистически обработать эти пробы и сравнить результаты с теми, которые уже получены на основе выборок из речи Симонова и Шолохова. Задача была бы решена.

Конечно, условия, в которых приходится решать задачи литературной атрибуции, могут оказаться неизмеримо сложнее, чем это только что изображено. Естественно, соответствующим образом усложнится и решение задачи. Особенно усложняется такое решение в случаях, когда много предполагаемых авторов безымянных текстов. В таких случаях пришлось бы изучать статистические характеристики речи многих авторов и со многими речевыми структурами сопоставлять речевую структуру безымянного произведения.

Конечно, может случиться и так, что предполагаемые авторы безымянного произведения не имеют отчетливо выраженного индивидуального стиля. Придется отыскивать тот структурный пласт речи, в котором каждый из авторов имеет все же свое лицо и оно может быть замечено и обрисовано с помощью статистики. Но общий принцип остается: прежде чем сравнивать особенности речи безымянного произведения с особенностями речи предполагаемого автора, нужно убедиться, что речь этого автора имеет улавливаемые статистикой особенности и нужно получить статистические данные об этих особенностях.

Вопросы атрибуции литературных текстов затрагивают и такое уникальное явление истории литературы, как «Слово о полку Игореве». Споры вокруг «Слова...» не утихают. Некоторые специалисты даже ставят под сомнение его подлинность (в который-то раз?). Почему бы не попытаться применить статистику для прояснения хотя бы части сложных и темных вопросов, рожденных историей «Слова...» и его изучения? Почему бы для начала не попробовать узнать, языку какого века — тринадцатого, шестнадцатого или восемнадцатого! — ближе язык «Слова...»? А ведь такая задача вполне разрешима, если предположить, что многие статистические закономерности языка и его функционирования существенно менялись от века XIII к веку XVI и от века XVI к веку XVIII.



6. Тут естествен переход к раздумьям о возможностях статистической методики в исследовании истории языка вообще и истории литературного языка в частности.

В конце концов, история языка имеет дело с изменением, развитием каких-то явлений языка и условиями этого развития — внутриязыковыми и внеязыковыми. Историка языка обычно не удовлетворяет сама по себе констатация таких-то и таких-то перемен языковых единиц и категорий: нужно объяснение этих перемен, т. е. нужно понимание их внутриязыковых и внеязыковых условий. Кроме того, становится все более ясным, что историк языка имеет дело и с изменениями в самом функционировании языка, его вариантов, его структурных областей, их единиц и категорий.

Понятно, что изучать изменения в функционировании языковой структуры без статистики, по-видимому, просто невозможно. Использование статистики резко расширяет горизонты достоверного знания о том, как изменяется работа языкового механизма, его отдельных узлов и элементов от эпохи к эпохе. Опыт соответствующего изучения некоторых явлений грамматического строя русского литературного языка XIX—XX вв. был показан мною в 1965 г. [1]. Разумеется, могут быть и иные подходы к изучению истории литературного языка с применением статистики. Важно сейчас представить другое, а именно то, что статистика позволяет оценивать факты, зафиксированные во многих и различных документах, позволяет получить обобщение языковых явлений, показанных большими текстовыми массивами. Поэтому при участии статистики лингвист может получить ответ и на такие вопросы, которых, обычно, применяя традиционные методики, он избегает. Например, как изменилась речевая активность частей речи русского литературного языка на протяжении XVIII—XX вв.? Какие изменения за то же время произошли в

[1] См.: Б. Н. Головин. Опыт вероятностно-статистического изучения некоторых явлений истории русского литературного языка XIX—XX вв. «Вопросы языкознания», 1965, № 3, стр. 137—146.

функционировании предложений различной структуры? Как менялась стилевая дифференциация структуры русского глагола в XVI—XVIII столетиях? Существует ли внутренняя зависимость между усилением речевой активности родительного падежа в XIX—XX вв. и функционированием предлогов? Можно ли предполагать, что развитие категории вида русского глагола в XIII—XVI вв. обусловлено влиянием характера глагольных основ? Устанавливается ли зависимость между развитием второго полногласия и падением редуцированных в XIII—XV вв.? Можно ли предполагать влияние классов и социальных групп на функционирование языка XIX—XX вв. и в чем оно могло выражаться? Как изменилось функционирование словарного состава русского языка от пушкинского периода к нашему времени? И т. д.

Уже говорилось о том, что статистика позволяет так организовать опыт, чтобы из некоторого ряда предполагаемых условий (или причин) наблюдаемого изменения выделить одно и проверить гипотезу о его воздействии на изменяющееся языковое явление.

Предположим, что развитие родительного определительного находится (или находилось) в зависимости от семантики грамматически ведущего или грамматически подчиненного члена именного сочетания. Разделим имена существительные на семантические группы в соответствии с тем, что дает интуитивное изучение языка. Получим текстовые выборки именных словосочетаний с родительным определительным. Установим, с какой частотой в совокупных выборках определенной длины встречаются имена намеченных семантических групп. Осуществим минимально-необходимую статистическую обработку данных. Если одни семантические группы дадут существенно более высокие частоты родительного определительного, чем другие семантические группы, значит, несомненна зависимость изучаемого явления от семантики имен; если к тому же будут осуществлены аналогичные наблюдения и на другом временном срезе языка, можно получить достоверную картину влияния семантики имен на развитие родительного определительного.

Проблема воздействия речевого творчества отдельных авторов на развитие тех или иных узлов и сторон языковой структуры может решаться в принципе подобным же образом. Пушкин — основоположник современного русского литературного языка. Это скорее аксиома, чем система доказательств. Интуиция поддерживает и защищает эту аксиому. И нет оснований сомневаться в ее истинности. Но наука нуждается и в том, чтобы получить детализированную картину изменений в функционировании русского литературного языка, вызванных прямым или косвенным воздействием речевого творчества Пушкина. Что именно и каким образом изменилось? Это можно установить. Нужно лишь последовательно, один за другим, сравнить структурные элементы нашего языка в пушкинском и послепушкинском, общелитературном, употреблении. Обнаружатся статистические схождения и расхождения, статистический аппарат позволит оценить их величину. Послепушкинские явления полезно сравнить с допушкинскими и также оценить с помощью статистики. Те явления в литературном языке послепушкинской поры, которые несущественно будут отличаться от пушкинских и существенно от предпушкинских, наблюдатель будет вправе признать оказавшимися под влиянием пушкинского речевого творчества.

Все это можно сделать. Можно установить и меру влияния на литературный язык (или его стилевые и иные варианты), оказанного и другими большими литераторами. Можно выделить традиции литературно-языкового развития, рожденные деятельностью представителей литературы, науки, политики, имевших национальное значение и получивших национальное признание.

Одним словом, очень многое можно сделать, если понять существо статистической методики. А оно состоит в том, что за наблюдаемыми частотами и долями в речи и языке, как правило, стоят закономерности, сама природа которых требует для ее понимания привлечения понятий, терминов и аппарата математической статистики. А так как этими закономерностями управляется не только функционирование, но и развитие языка, то статистическая методика оказывается незаменимым помощником лингвиста и тогда, когда он изучает явления языка в их современном состоянии, и тогда, когда он исследует явления языка в их историческом движении.

7. Неполно и бегло говоря о проблемах и перспективах статистического изучения языка и речи, нужно, видимо, сказать и о некоторых более сложных, чем описанные, вариантах статистической методики.

К их числу можно, прежде всего, отнести корреляционный анализ языковых фактов.

*Математическое понятие корреляции довольно сложно для филолога, не имеющего высшего математического образования.* Но, может быть, филологу и не обязательно становиться математиком каждый раз, когда приходится применять математические инструменты. Может быть, и в обсуждении возможностей и техники применения корреляционного анализа для решения лингвистических задач достаточно на первых порах знать, что корреляция — это связь, функциональная зависимость, существующая между двумя рядами явлений и устанавливаемая при помощи определенной статистической процедуры.

Методика корреляционного анализа оказывается полезной как раз тогда, когда ученый хочет проверить свою гипотезу о наличии зависимости между фактами ряда *А* и фактами ряда *Б*.

Так, лингвиста может интересовать вопрос о том, существует ли в действительности обсуждавшаяся многими учеными проблема «имя или глагол»? Иначе говоря, существует ли та антагонистическая зависимость между этими частями речи, которая как будто предполагается?

Для ответа на поставленные вопросы можно получить из некоторого числа выборок достаточного объема частоты имен существительных и глаголов и проверить гипотезу о зависимости между глагольным и именным частотными рядами. Эта зависимость может выразиться либо в том, что при увеличении частотности имен увеличивается и частотность глаголов, либо в том, что при увеличении частотности одной из двух частей речи частотность другой падает.

Но как же проверить нашу гипотезу? Как установить в кажущемся беспорядке частот ту или иную из двух предполагаемых тенденций или отсутствие и той и другой?

Для всего этого и потребуется техника, статистический аппарат корреляционного анализа.

Вот один из доступных лингвисту вариантов такой техники.

На основе данных, полученных в ряду выборок равного объема, составляется таблица, фиксирующая выборочные частоты, отклонения от средних частот, квадраты



отклонений и произведения отклонений выборочных частот двух явлений языка, между которыми предполагается зависимость, измеряемая при помощи корреляционного анализа.

| Выборки | $x$ | $a$ | $a^2$ | $y$ | $b$ | $b^2$ | $ab$ |
|---|---|---|---|---|---|---|---|
| 1-я | 49 | +4 | 16 | 15 | +1,5 | 2,25 | +6 |
| 2-я | 53 | +8 | 64 | 10 | —3,5 | 12,25 | —28 |
| 3-я | 58 | +13 | 169 | 6 | —7,5 | 56,25 | —97,5 |
| 4-я | 39 | —6 | 36 | 19 | +5,5 | 30,25 | —33 |
| 5-я | 37 | —8 | 64 | 16 | +2,5 | 6,25 | —20 |
| 6-я | 48 | +3 | 9 | 12 | —1,5 | 2,25 | —4,5 |
| 7-я | 43 | —2 | 4 | 12 | —1,5 | 2,25 | +3 |
| 8-я | 56 | +11 | 121 | 7 | —6,5 | 42,25 | —71,5 |
| 9-я | 38 | —7 | 49 | 17 | +3,5 | 12,25 | —24,5 |
| 10-я | 37 | —8 | 64 | 17 | +3,5 | 12,25 | —28 |
| 11-я | 39 | —6 | 36 | 18 | +4,5 | 20,25 | —27 |
| 12-я | 46 | +1 | 1 | 13 | —0,5 | 0,25 | —0,5 |
| 13-я | 48 | +3 | 9 | 14 | +0,5 | 0,25 | +1,5 |
| 14-я | 37 | —8 | 64 | 17 | +3,5 | 12,25 | —28 |
| 15-я | 45 | 0 | 0 | 11 | —2,5 | 6,25 | 0 |
| $\Sigma_{x_i}=673$ | | $\Sigma a_i^2=706$ | | $\Sigma_{y_i}=204$ | | $\Sigma b_i^2=217{,}75$ | |
| $\bar{x}=45$ | | $\bar{y}=13{,}5$ | | $\Sigma ab=-352$ | | | |

В нижней строке таблицы даны суммы всех выборочных частот каждого из двух сравниваемых явлений (в нашем примере $x$ — это частоты имен существительных, $y$ — частоты местоимений), суммы квадратов отклонений от средних частот каждого из двух сравниваемых явлений и сумма произведений двух отклонений от двух средних частот в одной и той же выборке. Таких табличных данных достаточно, чтобы вычислить так называемый коэффициент корреляции (он обычно обозначается буквой $r$), который как раз и должен показать наличие или отсутствие зависимости между двумя заинтересовавшими лингвиста явлениями. Вот формула для вычисления коэф-

фициента корреляции: $r = \frac{\Sigma ab}{\sqrt{\Sigma a_i^2 \cdot \Sigma b_i^2}}$. В формуле $a_i$ и $b_i$ — выборочные отклонения от средних выборочных частот $x$ и $y$; в произведении $a$ на $b$ надо сохранять знаки по правилам алгебраического умножения.

Коэффициент корреляции может иметь знак «плюс» или знак «минус» и может меняться по абсолютной величине от нуля до единицы. Знак «минус» при коэффициенте корреляции будет говорить наблюдателю о том, что между двумя явлениями есть отрицательная связь, т. е. такая, которая выражается в увеличении частотности одного из явлений при уменьшении частотности другого; знак «плюс» при коэффициенте корреляции известит наблюдателя о том, что между двумя явлениями существует положительная зависимость, т. е. такая, которая выражается в увеличении или уменьшении частотности одного явления при увеличении или соответственно уменьшении частотности другого. Чем больше абсолютная величина коэффициента корреляции, тем теснее связь между изучаемыми явлениями.

Таблица, которую читатель только что видел, содержит необходимые для вычисления коэффициента корреляции сведения, говорящие о частотах имен существительных и глаголов в научном литературоведческом тексте; длина каждой из 15 выборок — 100 знаменательных слов. Применим формулу для вычисления коэффициента корреляции $r = \frac{-352}{\sqrt{706 \cdot 238}} = -0{,}90$. Полученный коэффициент достаточно велик, чтобы признать наличие сильной отрицательной зависимости между именами существительными и местоимениями в процессе их функционирования в научно-публицистическом стиле (выборки были взяты из журнала «Вопросы литературы»). Этот коэффициент определенно указывает на то, что одной из закономерностей работы языкового механизма современного русского языка при обслуживании им нужд научной публицистики является отрицательное коррелирование имен существительных и местоимений, т. е. увеличение активности одной из этих частей речи за счет другой.

По наблюдениям Н. Барановской, в 30 выборках, взятых из авторской художественной речи А. И. Герцена и А. И. Гончарова (каждая выборка длиной в 500 знаменательных слов), оказалось такое количество имен существительных и местоимений (на первом месте — частота имен существительных, на втором — частота местоимений):

171—100, 187—98, 164—114, 186—85, 181—68, 168—98,
201—92, 150—109, 183—105, 158—106, 183—94, 169—109,
201—79, 202—76, 205—85, 170—102, 213—71, 190—96,
199—85, 218—73, 185—120, 196—92, 240—63, 170—83,
185—76, 184—94, 161—107, 164—106, 176—107, 162—101.

Обработка этих частот, осуществленная так, как показано было в таблице, дает: $\Sigma a_i^2 = 11\,684$, $\Sigma b_i^2 = 6120$, $\Sigma ab = -6351$. По формуле $r = \frac{-6351}{\sqrt{11\,684 \cdot 6120}} = -0{,}82$.

И этот опыт дал достаточно большой коэффициент корреляции, также с отрицательным знаком.

Но что значит «достаточно большой»?

Дело в том, что коэффициент корреляции полезен исследователю лишь тогда, когда он говорит о существовании зависимости между некоторыми явлениями в действительности, в самой «генеральной совокупности» изучаемых фактов, а не только в выборках из нее. Выборки могут давать некоторые величины коэффициента корреляции случайно, и коэффициент корреляции начинает обманывать наблюдателя, начинает сигнализировать наличие зависимости между элементами или явлениями, когда на самом деле никакой зависимости нет. Поэтому от наблюдателя требуется известная осторожность в использовании выборочных коэффициентов корреляции. Нужно знать те предельные величины этих коэффициентов, ниже которых они становятся сомнительными и не должны служить основанием для теоретических обобщений или практических выводов. Читателю может помочь таблица, построенная по графику в книге М. Езекиэла и К. Фокса «Методы анализа корреляций и регрессии» (М., 1966, стр. 309).

В двух описанных опытах изучения корреляционной зависимости между именами существительными и местоимениями были получены коэффициенты корреляции — 0,90 (при 15 наблюдениях) и — 0,82 (при 30 наблюдениях). Таблица позволит установить, что первому в 95 опытах из 100 (каждый опыт — 15 выборок по 100 знаменательных слов из однородного текста) соответствуют коэффициенты корреляции не менее 0,76, а в 5 опытах могут оказаться и мень-

**Минимальные величины истинных коэффициентов корреляции (при 5% вероятности меньшего их значения), соответствующие выборочным коэффициентам корреляции при различном числе наблюдений**

| Выборочный коэффициент | Минимальный истинный коэффициент при $K$ наблюдениях | | | | | |
|---|---|---|---|---|---|---|
| | $K=10$ | $K=15$ | $K=20$ | $K=30$ | $K=40$ | $K=50$ |
| 0,20 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| 0,25 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| 0,30 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| 0,35 | 0,00 | 0,00 | 0,00 | 0,03 | 0,10 | 0,14 |
| 0,40 | 0,00 | 0,00 | 0,00 | 0,12 | 0,16 | 0,19 |
| 0,45 | 0,00 | 0,00 | 0,08 | 0,18 | 0,22 | 0,25 |
| 0,50 | 0,00 | 0,04 | 0,16 | 0,24 | 0,28 | 0,30 |
| 0,55 | 0,00 | 0,16 | 0,23 | 0,30 | 0,34 | 0,37 |
| 0,60 | 0,04 | 0,24 | 0,30 | 0,37 | 0,40 | 0,43 |
| 0,65 | 0,18 | 0,31 | 0,37 | 0,44 | 0,47 | 0,49 |
| 0,70 | 0,28 | 0,39 | 0,45 | 0,51 | 0,54 | 0,56 |
| 0,75 | 0,37 | 0,48 | 0,53 | 0,58 | 0,61 | 0,63 |
| 0,80 | 0,48 | 0,57 | 0,61 | 0,66 | 0,68 | 0,70 |
| 0,85 | 0,60 | 0,64 | 0,70 | 0,74 | 0,76 | 0,77 |
| 0,90 | 0,71 | 0,76 | 0,80 | 0,83 | 0,84 | 0,85 |

шей величины. Второму полученному коэффициенту (0,82) в 95 опытах из 100 (один опыт — 30 выборок по 500 знаменательных слов) соответствуют коэффициенты корреляции не менее 0,69—0,70, а в 5 опытах могут иметь меньшую величину.

Полезно обратить внимание на то, что таблица хорошо показывает и обманывающие выборочные коэффициенты. Так, например, если мы при 20 наблюдениях получили выборочный коэффициент 0,35, мы не имеем права говорить о зависимости между изучаемыми явлениями: таблица говорит о том, что такому выборочному коэффициенту корреляции в более чем пяти опытах из ста может соответствовать нулевая зависимость между двумя рядами частот, т. е. отсутствие корреляции в действительности.

Наблюдателя языковых фактов, применяющего корреляционный анализ, подстерегает и еще одна исследовательская опасность — натолкнуться на достаточно большие коэффициенты мнимой, «бессмысленной» корреляции [1]. Жизнь сложна и многообразна, и совершенно случайно можно натолкнуться на два таких явления, возрастание и убывание частот которых во времени или пространстве имеет некоторое соответствие, наблюдаемый параллелизм, не говорящий, однако, ни о какой необходимой зависимости между явлениями. Поэтому от экспериментатора-исследователя, решившего применить корреляционный анализ, всегда требуется некоторая гипотеза о причинных или функциональных зависимостях между явлениями и умение интерпретировать полученные из опыта коэффициенты корреляции в духе этой гипотезы.

Но все эти призывы к осторожности не должны останавливать лингвиста в позиции постороннего наблюдателя усилий, осуществляемых в других науках. Можно и нужно смело испытывать еще один инструмент статистики — корреляционный анализ: развитие и функционирование языка дает для этого и основания, и поводы. Например, следует ли из грамматической зависимости имени прилагательного от имени существительного, что прилагательное подчинено существительному и в речевом функционировании, т. е. что частотность имени прилагательного растет и падает вместе с ростом и падением частотности имени существительного? И если такая зависимость существует, каковы ее основания, условия и причины? Одинакова ли она в разных языковых стилях и типах речи? Подобные же вопросы могут быть отнесены к имени и глаголу, глаголу и наречию, имени существительному и предлогу, глаголу и предлогу, имени и местоимению, местоимению и глаголу, предложениям простому и сложному, подлежащему и сказуемому, главным членам и членам второстепенным, имени существительному и второстепенным членам, глаголу и сказуемому и т. д.

Коэффициенты корреляции, полученные в двух описанных опытах, говорят убедительно об особой закономерности (или тенденции) функционального взаимодействия имен и местоимений: усиление активности имени существительного влечет за собой ослабление активности местоимений, и наоборот. Такую тенденцию можно назвать тенденцией функционального отталкиваний. По предвари-

[1] См. Дж. Эдни Юл и М. Дж. Кендэл. Теория статистики. М., 1960, стр. 263—265.

тельным данным, тенденцией функционального отталкивания связаны также имена существительные и глаголы (в опыте из 20 наблюдений получен $r=-0{,}54$), хотя есть основания думать, что эта тенденция испытывает большие колебания во времени и «речевом пространстве», т. е. в разных видах и типах речи. Противоположную тенденцию можно назвать тенденцией функционального притяжения. По предварительным же данным, она существует между именем существительным и именем прилагательным, между глаголом и предлогом (соответственно величины коэффициента корреляции равны $+0{,}59$ и $+0{,}51$ в опытах по 20 наблюдений в каждом); есть основания думать, что корреляционная зависимость между названными частями речи также имеет заметные колебания во времени и по языковым и речевым стилям. Но как бы ни были пока неуверенны и предварительны сведения о степени корреляционной зависимости между отдельными явлениями языка, сам факт такой зависимости устанавливается вполне определенно, что и позволяет с надеждой думать о будущем корреляционного анализа в науке о языке.

Читателю в связи со сказанным небесполезно узнать о том, что статистика, помимо коэффициента корреляции, знает величину $r^2$ (коэффициент корреляции, возведенный в квадрат), называемую коэффициентом детерминации. Он позволяет получить представление о той доле в изменениях частотного ряда явления $x$, которая вызывается влиянием изменений частотного ряда $y$. Так, если в опыте изучения функциональной зависимости имен существительных и местоимений были получены коэффициенты корреляции 0,90 и 0,82, то соответствующие коэффициенты детерминации будут равны 0,81 и 0,67. Это означает, что 81% (67%) изменений частотного ряда одной части речи обусловлены воздействием частотных изменений в ряду другой части речи. Правда, такой вывод не устраняет возможности общей внеструктурной основы, обусловливающей, в конечном счете, изменения и одного и другого частотного ряда. Но какова бы ни была природа функциональной зависимости между двумя частями речи (или иными явлениями языка), эта зависимость вскрывается и интенсивность ее измеряется при помощи коэффициента корреляции и коэффициента детерминации. Лингвисты едва ли останутся равнодушны к этим новым для них орудиям познания.



## ПОСЛЕСЛОВИЕ

Итак, автор завершает свою работу, посвященную объяснению необходимых лингвисту понятий и инструментов математической статистики и одного из вариантов их методического исследовательского применения в науке о языке. Но ведь современная лингвистическая статистика не сводится к одному этому варианту. Она представляет собой широкую — по разнообразию изучаемых языков и их структурных участков, по кругу конкретных теоретических и прикладных задач, по варьированию самой методики — область науки о языке. Изложение всех сведений о ней не было целью автора этой книги. Вместе с тем автор не в праве совершенно ничего не сказать читателю о современной лингвистической статистике в многообразии ее предметов, задач и методических разветвлений. Теперь, после того, как читатель получил некоторые сведения о самых необходимых лингвисту понятиях и инструментах математической статистики, автор может предложить читателю краткий обзор главных проблем и направлений современной — по преимуществу отечественной — лингвостатистики.

1. К настоящему времени вполне окрепло и уже дало науке и практике ряд самостоятельных исследовательских работ то направление, которое можно назвать *лексикографической статистикой*; это направление свою главную задачу видит в создании частотных словарей и разрабатывает связанные с этой задачей вопросы теории и методики. При создании частотных словарей различных языков перед составителями возникают десятки теоретических и методических вопросов. Прежде всего, составители словаря должны обеспечить достаточную надежность тех частотных показателей, которыми будут снабжены слова в лексиконе. Как эту надежность можно обеспечить? Какого объема выборку нужно взять? Очевидно, в общем виде на такой вопрос можно ответить: чем больше выборка, тем надежнее частотные показатели слов в словаре, т. е. тем больше словарь соответствует языковой реальности. Но практически выборка всегда будет меньше некоторого наилучшего максимума. Какой же объем выборки можно признать достаточным? И не зависит ли эта «достаточность» от того, как будет поставлена задача лекси-

кографом? И если лексикограф принял решение об объеме выборки, как будет именно эта выборка влиять на надежность частотных характеристик слов, принадлежащих к различным пластам лексики по активности речевого применения? Возникает и вторая не менее сложная задача — как оптимальным образом сформировать выборку, если словарь должен отражать функционирование лексики языка в целом, а не отдельных его стилей? Нужно ли брать из каждого стиля подвыборки одинакового объема или же следует каким-то образом учесть удельный вес каждого из функциональных стилей в языке определенного исторического периода? Допустимо ли и в каком соотношении включение в совокупную выборку лексического материала, взятого из текстов различных исторических периодов, т. е. какими хронологическими рамками должна быть ограничена выборка? Подобные вопросы возникнут и при составлении словаря, отражающего функционирование лексики отдельного функционального стиля. Эти задачи и вопросы даны лишь как иллюстрация тех трудностей, с которыми встречаются лексикографы-статистики.

К настоящему времени создано более трехсот частотных словарей и частотных списков слов различного объема, назначения и содержания. Первым был частотный словарь немецкого языка, подготовленный еще в конце XIX в. Кедингом на основе обследования текстов, включавших около одиннадцати миллионов словоупотреблений <106> [1]. Известны частотные словари — английского языка Торндайка и Лоджа <107>, французского языка Вандер Беке <109>, испанского языка Гарсиа Оса <95>, четырехязычный частотный словарь Элен Итон <94>, русского языка Йоссельсона <103>, разговорной русской речи Вакара <108>, чешского языка Елинка, Бечки и Тешителовой <102>, испанского и румынского языков Жуильяна <104, 105> и другие.

В Советском Союзе первым частотным словарем стал, по существу, «Словарь языка Пушкина», хотя в нем частотные признаки слов даны лишь как вспомогательные характеристики качественного описания лексики. Попытку решить чисто статистические задачи в лексикографическом отображении лексики современного русского языка впер-

[1] Цифра в угловых скобках обозначает номер названия работы в списке литературы, помещенном в конце книги.



вые были сделаны коллективом под руководством Э. А. Штейнфельдт <77>. В Риге в 1966 году увидел свет первый том частотного словаря латышского языка <100>. Коллективом лингвистов и математиков Ленинградского и Горьковского университетов под руководством Л. Н. Засориной ведется подготовка большого частотного словаря современного русского литературного языка <22>. В 1968 году Университетом дружбы народов имени Патриса Лумумбы опубликованы частотные списки наиболее употребительных слов русской разговорной речи <21>. Создаются частотные словари и списки слов, ориентированные на отраслевые разновидности научного и производственно-технического стилей («подъязыков»); в организации и осуществлении такого рода работ в Советском Союзе видное место принадлежит Р. Г. Пиотровскому и возглавляемой им группе ученых «Статистика речи» <64>. В различного рода статьях и иных публикациях активно обсуждаются вопросы теории и методики составления частотных словарей; здесь могут быть названы имена Р. М. Фрумкиной, Л. Н. Засориной, П. М. Алексеева, В. М. Андрющенко, В. М. Калинина, В. А. Московича и других <3, 5, 22, 26, 51, 70, 71, 72>. Вместе с развитием статистической лексикографии все чаще осуществляются опыты применения статистики в изучении вопросов лексикологии.

2. Особый круг проблем и задач связан с той областью современного языкознания, которую нередко называют стилостатистикой, имея в виду применение статистических методик в изучении языковых и речевых стилей. Эта область науки, как и статистическая лексикография, уже представлена многими именами и конкретными исследовательскими работами. В сущности, пионером в стилостатистике в России был известный народоволец, узник Шлиссельбурга, Н. А. Морозов <52>. И хотя первые опыты в применении статистики к изучению стилей были не вполне удачны и вызвали критическую оценку специалиста-математика <45>, сама исследовательская идея Н. А. Морозова не пропала. В настоящее время и в Советском Союзе, и за его рубежами выполнены или выполняются многие конкретные исследования языковых и речевых стилей. В результате созрело понимание того, что функционирование языка вариативно, и эта вариативность, обусловленная различием видов человеческой деятельности (а также и иными условиями), лежит в основе стиле-

вой дифференциации языка и речи. Проблема изучения стилей стала осознаваться по-иному. По сравнению с недавним прошлым, она приобрела, под влиянием статистического изучения фактов, большую отчетливость и определенность, и само понимание языковых и речевых стилей становится в чем-то иным <19, 20>. Проблема стилевой дифференциации средств языка все отчетливее осознается как одна из главных проблем современной науки о языке.

Интересно и интенсивно применяют статистическую методику в изучении языковых и речевых стилей ученые Украины: коллектив, который возглавляет В. И. Перебийнос, уже получил богатые результаты, позволяющие с большой надеждой смотреть на будущее этих исследований <65>. Нельзя не упомянуть в этом беглом обзоре и преподавателей Саратовского государственного университета, усилиями которых получены новые результаты, обогащающие научные представления о стилевой дифференциации современного русского литературного языка, в особенности в области синтаксиса <11>. Около десяти лет продолжается эксперимент по статистическому изучению стилевой дифференциации грамматического строя русского литературного языка XIX—XX вв., осуществляемый в Горьковском государственном университете студентами, аспирантами и преподавателями; результаты этого эксперимента фактически еще не опубликованы <17, 20>. Новые материалы, относящиеся к решению проблемы стилей языка и стилей речи, обнародованы Г. А. Лесскисом и С. И. Кауфманом <28, 29, 30, 42, 43, 44>. Большой известностью пользуются в кругу специалистов работы П. Гиро, посвященные обоснованию статистического изучения стилевого функционирования лексики <96>. Не так давно высказал свои интересные соображения о применении математико-статистических методов в стилистике чехословацкий ученый Й. Мистрик <47>. Перечень проблем, тем, имен, изданий, связанных со стилостатистикой, можно без особых поисков и усилий продолжить.

3. Внимание исследователей привлекают и общие вопросы статистического изучения языка, квантитативного подхода к языковым структурам. В этом направлении многое сделано Н. Д. Андреевым и его учениками <4, 63>. Методический алгоритм статистико-комбинаторного анализа, предложенный Н. Д. Андреевым, представляет опре-



деленный интерес для всех, кто ищет более рациональные пути и методики статистического исследования единиц и категорий языка на разных уровнях его структуры. Новые аспекты количественного изучения языка ищет и находит уже упоминавшаяся группа «Статистика речи», руководимая профессором Р. Г. Пиотровским. Не только стилостатистические, но и лингво-статистические интересы имеют киевская и горьковская группы исследователей. Осуществлены обнадеживающие опыты статистических подходов к изучению семантических полей В. А. Московичем и А. Я. Шайкевичем <49, 50>, к изучению значений глаголов и глагольного управления Ю. Д. Апресяном <6>, изучению объективной стилевой принадлежности текстов — А. Я. Шайкевичем <75> и т. д. Получены неоднократно статистические показатели функционирования фонем в разных языках. Хэрданом предприняты попытки разработать общую теорию квантитативного осмысления языка <97, 98, 99>. Появляются все новые и новые работы, авторы которых обсуждают место количественных аспектов изучения языка в общей системе современного языкознания <1, 23, 39, 59, 69>.

4. Все отчетливее становятся видимы практические результаты лингво-статистических исследований — в построении систем научно-технической информационной службы, в автоматизации части работ, связанных с изучением текстов, в переводах с одного языка на другой, в преподавании иностранных и родных языков, в телефонной и радиосвязи и т. д. Появляются исследования, ориентированные на статистическое решение разнообразных прикладных задач. О поисках в этой области можно в какой-то мере судить по материалам конференций <46, 73>, публикациям группы «Статистика речи», статьям В. А. Московича и других ученых <3, 8, 26, 51, 63>.

5. Ширится изучение стиха при участии статистики. Уже получены интересные результаты в описании и осмыслении ритма, размеры, рифмы русского стиха <9, 10, 13, 14, 33, 34, 35, 36, 37, 38>. Особенно примечательны усилия в этой области М. Л. Гаспарова, А. М. Кондратова, С. Боброва. Большую помощь лингвистам и филологам-литературоведам в статистическом изучении стиха оказывает А. Н. Колмогоров. Осуществляются первые опыты систематического изучения ритма художественной

прозы <24>. Первые попытки применения статистики в описании внешней стороны художественной речи, предпринятые в 20-е годы <56, 62, 74>, не были бесплодными. В настоящее время поэтика все шире и смелее пользуется подходами, понятиями, методическими инструментами, связанными с математической статистикой и теорией вероятности.

6. Известны выдающиеся успехи советских лингвистов в расшифровке древних текстов. Эти успехи во многом зависят от применения особых методик, использующих понятия комбинаторной статистики. Ю. В. Кнорозов создал оригинальную методику, позволившую расшифровать, с помощью статистики, большую часть письменных документов народа майя <31>. В. В. Шеворошкин осуществил серию успешных опытов по дешифровке карийских надписей на основе оригинальной методики — также с участием статистики <76>. Б. В. Сухотин создал свою теоретическую и методическую концепцию лингвистической дешифровки, привлекшую внимание специалистов строгостью исходных позиций и решений; и в этой концепции важное место отведено статистике <68>. Таким образом, можно говорить о возникновении в системе лингвистической и исторической науки особого направления, представители которого развивают теорию и практику дешифровки письменных текстов на базе статистических понятий.

7. В общем кратком обзоре современной отечественной лингвистической статистики (или, может быть, лучше — статистической лингвистики) нельзя не упомянуть о весьма интересных и поучительных опытах измерений информационной «емкости» знаков языка на различных уровнях его структуры. Подход к структуре языка и речевых цепей с позиций современной теории информации обещает много нового. И хотя необходимые для осуществления такого подхода методики еще не вполне надежны, они все же позволили получить совершенно новые количественные оценки единиц языка в разных типах текстов. Особо интересны и целенаправленны эксперименты в области информационного измерения языка, осуществляемые в группе Р. Г. Пиотровского, а также по методике А. Н. Колмогорова.

Статистическая лингвистика — это один из вариантов современного языкознания; у нее свои задачи, свои методики, свои заботы и трудности, свое содержательное настоящее и интересное будущее.

Можно и нужно с надеждой думать о том времени, когда статистика органически войдет в мышление и исследовательскую практику каждого лингвиста, а возможно — и каждого филолога. Описания и истолкования языка и речи, языковых и речевых стилей, функционирования языка и его развития получат мощное усиление и обновление на основе органического синтеза качественных и количественных сведений и представлений — в полном соответствии с той реальностью, которую все мы привычно обозначаем словами «язык» и «речь». И если в этой реальности количественные показатели и количественные закономерности занимают заметное место (а сомневаться теперь в этом не приходится), наука о языке должна с этим считаться, и она делает это. И чем отчетливее будут представлять себе лингвисты количественные аспекты языка и речи, связанные с этими аспектами проблемы, тематику и методики, тем успешнее лингвистика будет решать свои теоретические и прикладные задачи.

*Приложение 1*

**Квадраты целых и дробных чисел от единицы до десяти**

| Число | Его квадрат | Число | Его квадрат |
|---|---|---|---|
| 1,00 | 1,00 | 2,35 | 5,52 |
| 1,05 | 1,10 | 2,40 | 5,76 |
| 1,10 | 1,21 | 2,45 | 6,00 |
| 1,15 | 1,32 | 2,50 | 6,25 |
| 1,20 | 1,44 | 2,55 | 6,50 |
| 1,25 | 1,56 | 2,60 | 6,76 |
| 1,30 | 1,69 | 2,65 | 7,02 |
| 1,35 | 1,82 | 2,70 | 7,29 |
| 1,40 | 1,96 | 2,75 | 7,56 |
| 1,45 | 2,10 | 2,80 | 7,84 |
| 1,50 | 2,25 | 2,85 | 8,12 |
| 1,55 | 2,40 | 2,90 | 8,41 |
| 1,60 | 2,56 | 2,95 | 8,70 |
| 1,65 | 2,72 | 3,00 | 9,00 |
| 1,70 | 2,89 | 3,05 | 9,30 |
| 1,75 | 3,06 | 3,10 | 9,61 |
| 1,80 | 3,24 | 3,15 | 9,92 |
| 1,85 | 3,42 | 3,20 | 10,24 |
| 1,90 | 3,61 | 3,25 | 10,56 |
| 1,95 | 3,80 | 3,30 | 10,89 |
| 2,00 | 4,00 | 3,35 | 11,22 |
| 2,05 | 4,20 | 3,40 | 11,56 |
| 2,10 | 4,41 | 3,45 | 11,90 |
| 2,15 | 4,62 | 3,50 | 12,25 |
| 2,20 | 4,84 | 3,55 | 12,60 |
| 2,25 | 5,06 | 3,60 | 12,96 |
| 2,30 | 5,29 | 3,65 | 13,32 |



| Число | Его квадрат | Число | Его квадрат |
|---|---|---|---|
| 3,70 | 13,69 | 5,65 | 31,92 |
| 3,75 | 14,06 | 5,70 | 32,49 |
| 3,80 | 14,44 | 5,75 | 33,06 |
| 3,85 | 14,82 | 5,80 | 33,64 |
| 3,90 | 15,21 | 5,85 | 34,22 |
| 3,95 | 15,60 | 5,90 | 34,81 |
| 4,00 | 16,00 | 5,95 | 35,40 |
| 4,05 | 16,40 | 6,00 | 36,00 |
| 4,10 | 16,81 | 6,05 | 36,60 |
| 4,15 | 17,22 | 6,10 | 37,21 |
| 4,20 | 17,64 | 6,15 | 37,82 |
| 4,25 | 18,06 | 6,20 | 38,44 |
| 4,30 | 18,49 | 6,25 | 39,06 |
| 4,35 | 18,92 | 6,30 | 39,69 |
| 4,40 | 19,36 | 6,35 | 40,32 |
| 4,45 | 19,80 | 6,40 | 40,96 |
| 4,50 | 20,25 | 6,45 | 41,60 |
| 4,55 | 20,70 | 6,50 | 42,25 |
| 4,60 | 21,16 | 6,55 | 42,90 |
| 4,65 | 21,62 | 6,60 | 43,56 |
| 4,70 | 22,09 | 6,65 | 44,22 |
| 4,75 | 22,56 | 6,70 | 44,89 |
| 4,80 | 23,04 | 6,75 | 45,56 |
| 4,85 | 23,52 | 6,80 | 46,24 |
| 4,90 | 24,01 | 6,85 | 46,92 |
| 4,95 | 24,50 | 6,90 | 47,61 |
| 5,00 | 25,00 | 6,95 | 48,30 |
| 5,05 | 25,50 | 7,00 | 49,00 |
| 5,10 | 26,01 | 7,05 | 49,70 |
| 5,15 | 26,52 | 7,10 | 50,41 |
| 5,20 | 27,04 | 7,15 | 51,12 |
| 5,25 | 27,56 | 7,20 | 51,84 |
| 5,30 | 28,09 | 7,25 | 52,56 |
| 5,35 | 28,62 | 7,30 | 53,29 |
| 5,40 | 29,16 | 7,35 | 54,02 |
| 5,45 | 29,70 | 7,40 | 54,76 |
| 5,50 | 30,25 | 7,45 | 55,50 |
| 5,55 | 30,80 | 7,50 | 56,25 |
| 5,60 | 31,36 | 7,55 | 57,00 |

| Число | Его квадрат | Число | Его квадрат |
|---|---|---|---|
| 7,60 | 57,76 | 8,85 | 78,32 |
| 7,65 | 58,52 | 8,90 | 79,21 |
| 7,70 | 59,29 | 8,95 | 80,10 |
| 7,75 | 60,06 | 9,00 | 81,00 |
| 7,80 | 60,84 | 9,05 | 81,90 |
| 7,85 | 61,62 | 9,10 | 82,81 |
| 7,90 | 62,41 | 9,15 | 83,72 |
| 7,95 | 63,20 | 9,20 | 84,64 |
| 8,00 | 64,00 | 9,25 | 85,56 |
| 8,05 | 64,80 | 9,30 | 86,49 |
| 8,10 | 65,61 | 9,35 | 87,42 |
| 8,15 | 66,42 | 9,40 | 88,36 |
| 8,20 | 67,24 | 9,45 | 89,30 |
| 8,25 | 68,06 | 9,50 | 90,25 |
| 8,30 | 68,89 | 9,55 | 91,20 |
| 8,35 | 69,72 | 9,60 | 92,16 |
| 8,40 | 70,56 | 9,65 | 93,12 |
| 8,45 | 71,40 | 9,70 | 94,09 |
| 8,50 | 72,25 | 9,75 | 95,06 |
| 8,55 | 73,10 | 9,80 | 96,04 |
| 8,60 | 73,96 | 9,85 | 97,02 |
| 8,65 | 74,82 | 9,90 | 98,01 |
| 8,70 | 75,69 | 9,95 | 99,00 |
| 8,75 | 76,56 | 10,00 | 100,00 |
| 8,80 | 77,44 | | |

Читатель, конечно, помнит, что при переносе в числе запятой на один знак вправо или влево в квадрате числа запятая переносится в ту же сторону на два знака.

*Приложение 2*

**Квадратные корни из целых и дробных чисел от единицы до ста**

| Число | Квадратный корень | Число | Квадратный корень |
|---|---|---|---|
| 1,0 | 1,00 | 3,7 | 1,92 |
| 1,1 | 1,05 | 3,8 | 1,95 |
| 1,2 | 1,10 | 3,9 | 1,97 |
| 1,3 | 1,14 | 4,0 | 2,00 |
| 1,4 | 1,18 | 4,1 | 2,02 |
| 1,5 | 1,22 | 4,2 | 2,05 |
| 1,6 | 1,26 | 4,3 | 2,07 |
| 1,7 | 1,30 | 4,4 | 2,10 |
| 1,8 | 1,34 | 4,5 | 2,12 |
| 1,9 | 1,38 | 4,6 | 2,14 |
| 2,0 | 1,41 | 4,7 | 2,17 |
| 2,1 | 1,45 | 4,8 | 2,19 |
| 2,2 | 1,48 | 4,9 | 2,21 |
| 2,3 | 1,52 | 5,0 | 2,24 |
| 2,4 | 1,55 | 5,1 | 2,26 |
| 2,5 | 1,58 | 5,2 | 2,28 |
| 2,6 | 1,61 | 5,3 | 2,30 |
| 2,7 | 1,64 | 5,4 | 2,32 |
| 2,8 | 1,67 | 5,5 | 2,35 |
| 2,9 | 1,70 | 5,6 | 2,37 |
| 3,0 | 1,73 | 5,7 | 2,39 |
| 3,1 | 1,76 | 5,8 | 2,41 |
| 3,2 | 1,79 | 5,9 | 2,43 |
| 3,3 | 1,82 | 6,0 | 2,45 |
| 3,4 | 1,84 | 6,1 | 2,47 |
| 3,5 | 1,87 | 6,2 | 2,49 |
| 3,6 | 1,90 | 6,3 | 2,51 |

| Число | Квадратный корень | Число | Квадратный корень |
|---|---|---|---|
| 6,4 | 2,53 | 11,5 | 3,39 |
| 6,5 | 2,55 | 12,0 | 3,46 |
| 6,6 | 2,57 | 12,5 | 3,54 |
| 6,7 | 2,59 | 13,0 | 3,61 |
| 6,8 | 2,61 | 13,5 | 3,67 |
| 6,9 | 2,63 | 14,0 | 3,74 |
| 7,0 | 2,65 | 14,5 | 3,81 |
| 7,1 | 2,66 | 15,0 | 3,87 |
| 7,2 | 2,68 | 15,5 | 3,94 |
| 7,3 | 2,70 | 16,0 | 4,00 |
| 7,4 | 2,72 | 16,5 | 4,06 |
| 7,5 | 2,74 | 17,0 | 4,12 |
| 7,6 | 2,76 | 17,5 | 4,18 |
| 7,7 | 2,77 | 18,0 | 4,24 |
| 7,8 | 2,79 | 18,5 | 4,30 |
| 7,9 | 2,81 | 19,0 | 4,36 |
| 8,0 | 2,83 | 19,5 | 4,42 |
| 8,1 | 2,85 | 20,0 | 4,47 |
| 8,2 | 2,86 | 20,5 | 4,53 |
| 8,3 | 2,88 | 21,0 | 4,58 |
| 8,4 | 2,90 | 21,5 | 4,64 |
| 8,5 | 2,92 | 22,0 | 4,69 |
| 8,6 | 2,93 | 22,5 | 4,74 |
| 8,7 | 2,95 | 23,0 | 4,80 |
| 8,8 | 2,97 | 23,5 | 4,85 |
| 8,9 | 2,98 | 24,0 | 4,90 |
| 9,0 | 3,00 | 24,5 | 4,95 |
| 9,1 | 3,02 | 25,0 | 5,00 |
| 9,2 | 3,03 | 25,5 | 5,05 |
| 9,3 | 3,05 | 26,0 | 5,10 |
| 9,4 | 3,07 | 26,5 | 5,15 |
| 9,5 | 3,08 | 27,0 | 5,20 |
| 9,6 | 3,10 | 27,5 | 5,24 |
| 9,7 | 3,11 | 28,0 | 5,29 |
| 9,8 | 3,13 | 28,5 | 5,34 |
| 9,9 | 3,15 | 29,0 | 5,39 |
| 10,0 | 3,16 | 29,5 | 5,43 |
| 10,5 | 3,24 | 30,0 | 5,48 |
| 11,0 | 3,32 | 30,5 | 5,52 |



| Число | Квадратный корень | Число | Квадратный корень |
|---|---|---|---|
| 31,0 | 5,57 | 50,5 | 7,11 |
| 31,5 | 5,61 | 51,0 | 7,14 |
| 32,0 | 5,66 | 51,5 | 7,18 |
| 32,5 | 5,70 | 52,0 | 7,21 |
| 33,0 | 5,74 | 52,5 | 7,25 |
| 33,5 | 5,79 | 53,0 | 7,28 |
| 34,0 | 5,83 | 53,5 | 7,31 |
| 34,5 | 5,87 | 54,0 | 7,35 |
| 35,0 | 5,92 | 54,5 | 7,38 |
| 35,5 | 5,96 | 55,0 | 7,42 |
| 36,0 | 6,00 | 55,5 | 7,45 |
| 36,5 | 6,04 | 56,0 | 7,48 |
| 37,0 | 6,08 | 56,5 | 7,52 |
| 37,5 | 6,12 | 57,0 | 7,55 |
| 38,0 | 6,16 | 57,5 | 7,58 |
| 38,5 | 6,20 | 58,0 | 7,62 |
| 39,0 | 6,24 | 58,5 | 7,65 |
| 39,5 | 6,28 | 59,0 | 7,68 |
| 40,0 | 6,32 | 59,5 | 7,71 |
| 40,5 | 6,36 | 60,0 | 7,75 |
| 41,0 | 6,40 | 60,5 | 7,78 |
| 41,5 | 6,44 | 61,0 | 7,81 |
| 42,0 | 6,48 | 61,5 | 7,84 |
| 42,5 | 6,52 | 62,0 | 7,87 |
| 43,0 | 6,56 | 62,5 | 7,91 |
| 43,5 | 6,60 | 63,0 | 7,94 |
| 44,0 | 6,63 | 63,5 | 7,97 |
| 44,5 | 6,67 | 64,0 | 8,00 |
| 45,0 | 6,71 | 64,5 | 8,03 |
| 45,5 | 6,75 | 65,0 | 8,06 |
| 46,0 | 6,78 | 65,5 | 8,09 |
| 46,5 | 6,82 | 66,0 | 8,12 |
| 47,0 | 6,86 | 66,5 | 8,15 |
| 47,5 | 6,89 | 67,0 | 8,19 |
| 48,0 | 6,93 | 67,5 | 8,22 |
| 48,5 | 6,96 | 68,0 | 8,25 |
| 49,0 | 7,00 | 68,5 | 8,28 |
| 49,5 | 7,04 | 69,0 | 8,31 |
| 50,0 | 7,07 | 69,5 | 8,34 |

| Число | Квадратный корень | Число | Квадратный корень |
|---|---|---|---|
| 70,0 | 8,37 | 85,5 | 9,25 |
| 70,5 | 8,40 | 86,0 | 9,27 |
| 71,0 | 8,43 | 86,5 | 9,30 |
| 71,5 | 8,46 | 87,0 | 9,33 |
| 72,0 | 8,49 | 87,5 | 9,35 |
| 72,5 | 8,51 | 88,0 | 9,38 |
| 73,0 | 8,54 | 88,5 | 9,41 |
| 73,5 | 8,57 | 89,0 | 9,43 |
| 74,0 | 8,60 | 89,5 | 9,46 |
| 74,5 | 8,63 | 90,0 | 9,49 |
| 75,0 | 8,66 | 90,5 | 9,51 |
| 75,5 | 8,69 | 91,0 | 9,54 |
| 76,0 | 8,72 | 91,5 | 9,57 |
| 76,5 | 8,75 | 92,0 | 9,59 |
| 77,0 | 8,77 | 92,5 | 9,62 |
| 77,5 | 8,80 | 93,0 | 9,64 |
| 78,0 | 8,83 | 93,5 | 9,67 |
| 78,5 | 8,86 | 94,0 | 9,70 |
| 79,0 | 8,89 | 94,5 | 9,72 |
| 79,5 | 8,92 | 95,0 | 9,75 |
| 80,0 | 8,94 | 95,5 | 9,77 |
| 80,5 | 8,97 | 96,0 | 9,80 |
| 81,0 | 9,00 | 96,5 | 9,82 |
| 81,5 | 9,03 | 97,0 | 9,85 |
| 82,0 | 9,06 | 97,5 | 9,87 |
| 82,5 | 9,08 | 98,0 | 9,90 |
| 83,0 | 9,11 | 98,5 | 9,92 |
| 83,5 | 9,14 | 99,0 | 9,95 |
| 84,0 | 9,17 | 99,5 | 9,97 |
| 84,5 | 9,19 | 100,0 | 10,00 |
| 85,0 | 9,22 | | |

Квадратные корни из чисел, находящихся в интервалах между числами таблицы, лингвист может взять приближенно. При переносе запятой в числе на два знака вправо или влево запятая в квадратном корне переносится в ту же сторону на один знак.

## ОСНОВНАЯ ЛИТЕРАТУРА


1. Адмони В. Г. Качественный и количественный анализ грамматических явлений. «Вопросы языкознания», 1963, № 4.

2. Адмони В. Г. Размер предложения и словосочетания как явление синтаксического строя. «Вопросы языкознания», 1966, № 4.

3. Алексеев П. М. Частотный словарь английского подъязыка электроники. Автореферат кандидатской диссертации. Л., 1965.

4. Андреев Н. Д. Статистико-комбинаторные методы в теоретическом и прикладном языкознании. Л., «Наука», 1967.

5. Андрющенко В. М. Новые работы в области статистической лексикографии. «Вопросы языкознания», 1968, № 5.

6. Апресян Ю. Д. О сильном и слабом управлении. «Вопросы языкознания», 1964, № 3.

7. Ахманова О. С., Мельчук И. А., Падучева Е. В., Фрумкина Р. М. О точных методах исследования языка. Изд-во МГУ, 1961.

8. Белоногов Г. Г. О некоторых статистических закономерностях в русской письменной речи. «Вопросы языкознания», 1962, № 1.

9. Бобров С. Теснота стихотворного ряда. «Русская литература», 1965, № 3.

10. Бобров С. Русский тонический стих с ритмом неопределенной четности и варьируемой силлабикой. «Русская литература», 1967, № 1.

11. «Вопросы стилистики», вып. 3, под ред. О. Б. Сиротининой. Изд-во Саратовского государственного университета, 1969.

12. «Вопросы статистики речи» (Материалы совещания). Изд-во ЛГУ, 1958.

13. Гаспаров М. Л. Вольный хорей и вольный ямб Маяковского. «Вопросы языкознания», 1965, № 1.

14. Гаспаров М. Л. Ямб и хорей советских поэтов и проблема эволюции русского стиха. «Вопросы языкознания», 1967, № 3.

15. Головин Б. Н. О вероятностно-статистическом изучении стилевой дифференциации языка. Семинар «Автоматизация инфор-

мационных работ и вопросы математической лингвистики». Киев, 1964.

16. Головин Б. Н. Опыт вероятностно-статистического изучения некоторых явлений истории русского литературного языка XIX—XX вв. «Вопросы языкознания», 1965, № 3.

17. Головин Б. Н. Из курса лекций по лингвистической статистике. Горький, 1966.

18. Головин Б. Н. К вопросу о вероятностно-статистическом понимании стиля языка и стиля речи. В кн.: «Ученые записки (НИИ ПМК ГГУ). Прикладная математика и кибернетика». Горький. 1967.

19. Головин Б. Н., Урамбашев И. В. О статистических признаках стилевой дифференциации глагольных форм современного русского литературного языка. В кн.: «Ученые записки (НИИ ПМК ГГУ). Прикладная математика и кибернетика». Горький, 1967.

20. Головин Б. Н. О стилях языка и их изучении. «Русский язык в школе», 1968, № 4.

21. «2380 слов, наиболее употребительных в русской разговорной речи». М., Изд-во Университета дружбы народов им. П. Лумумбы, 1968.

22. Засорина Л. Н. Автоматизация и статистика в лексикографии. Изд-во ЛГУ, 1966.

23. Зиндер Л. Р., Строева Т. В. К вопросу о применении статистики в языкознании. «Вопросы языкознания», 1968, № 6.

24. Иванова Г. Н. Ритмика русской прозы. Автореферат кандидатской диссертации. М., 1969.

25. Йоссельсон Г. Г. Подсчет слов и частотный анализ грамматических категорий русского литературного языка. В кн.: «Автоматизация в лингвистике». «Наука». М. — Л., 1966.

26. Калинин В. М. О статистике литературного текста. «Вопросы языкознания», 1968, № 6.

27. Калинин В. М. Некоторые статистические законы математической лингвистики. «Проблемы кибернетики», вып. II. М., 1964.

28. Кауфман С. И. Об именном характере технического стиля. (На материале английской литературы.) «Вопросы языкознания», 1961, № 5.

29. Кауфман С. И. Количественный анализ общеязыковых категорий, определяющих качественные особенности стиля. В сб.: «Вопросы романо-германского языкознания». Коломна, 1961.

30. Кауфман С. И. Выражение логической последовательности в техническом стиле. В сб.: «Вопросы романо-германского языкознания». Коломна, 1961.



31. Кнорозов Ю. В. Система письма древних майя. М., «Наука», 1965.

32. Колмогоров А. Н., Кондратов А. М. Ритмика поэм Маяковского. «Вопросы языкознания», 1962, № 3.

33. Колмогоров А. Н. К изучению ритмики Маяковского. «Вопросы языкознания», 1963, № 4.

34. Колмогоров А. Н., Прохоров А. В. О дольнике современной русской поэзии. «Вопросы языкознания», 1963, № 6.

35. Колмогоров А. Н., Прохоров А. В. О дольнике современной русской поэзии. «Вопросы языкознания», 1964, № 1.

36. Колмогоров А. Н. Замечания по поводу анализа ритма «Стихов о советском паспорте» Маяковского. «Вопросы языкознания», 1965, № 3.

37. Кондратов А. М. Эволюция ритмики В. В. Маяковского. «Вопросы языкознания», 1962, № 5.

38. Кондратов А. М. Статистика типов русской рифмы. «Вопросы языкознания», 1963, № 6.

39. Конюс Е. А. Опыт применения статистического метода в области исторической лексикологии. «Вопросы языкознания», 1964, № 2.

40. Котелова Н. З. О применении объективных и точных критериев описания сочетаемости слов. «Вопросы языкознания», 1965, № 4.

41. Левин Ю. И. О количественных характеристиках распределения символов в тексте. «Вопросы языкознания», 1967, № 6.

42. Лесскис Г. А. О размере предложений в русской научной и художественной прозе 60-х гг. XIX в. «Вопросы языкознания», 1962, № 2.

43. Лесскис Г. А. О зависимости между размером предложения и характером текста. «Вопросы языкознания», 1963, № 3.

44. Лесскис Г. А. О зависимости между размером предложения и его структурой в разных видах текста. «Вопросы языкознания», 1964, № 3.

45. Марков А. А. Об одном применении статистического метода. «Известия АН (Bulletin)», т. 10, серия VI, № 4. П., 1915.

46. «Межвузовская конференция по вопросам частотных словарей и автоматизации лингвистических работ. Тезисы докладов и сообщений». Изд-во ЛГУ, 1966.

47. Мистрик И. Математико-статистические методы в стилистике. «Вопросы языкознания», 1967, № 3.

48. Москович В. А. Глубина и длина слов в естественных языках. «Вопросы языкознания», 1967, № 6.

49. Москович В. А. Статистика и семантика. М., «Наука», 1969.

50. Москович В. А. Опыт квантитативной типологии семантического поля. «Вопросы языкознания», 1965, № 4.

51. Москович В. А. Автоматизация некоторых аспектов лингвистической работы. «Вопросы языкознания», 1966, № 1.

52. Морозов Н. А. Лингвистические спектры. «Известия АН. Отделение русского языка и словесности», кн. 1—4, т. XX, 1915.

53. Никонов В. А. Статистика падежей русского языка. «Машинный перевод и прикладная лингвистика», вып. 3 (10). М., 1959.

54. Никонов В. А. Личные имена в современной России. «Вопросы языкознания», 1967, № 6.

55. Пап Ф. Количественный анализ словарной структуры некоторых русских текстов. «Вопросы языкознания», 1961, № 6.

56. Петерсон М. Н. Современный русский язык. М., 1929.

57. Петрова Н. В., Пиотровский Р. Г. Слово, контекст, морфология. «Вопросы языкознания», 1966, № 2.

58. Пиотровский Р. Г. О математическом языкознании. «Русский язык в национальной школе», 1961, № 2.

59. Пиотровский Р. Г. Математическое языкознание, его перспективы. «Вестник высшей школы», 1964, № 6.

60. Пиотровский Р. Г. Информационные измерения печатного текста. В кн.: «Энтропия языка и статистика речи». Минск, 1966.

61. Пиотровский Р. Г. Информационные измерения языка. Л., «Наука», 1968.

62. Пешковский А. М. 10.000 звуков русской речи. В кн.: «Методика родного языка, лингвистика, стилистика». Л. — М., 1925.

63. Статистико-комбинаторное моделирование языков», под ред. Н. Д. Андреева. М. — Л., «Наука», 1965.

64. «Статистика речи» (под ред. Р. Г. Пиотровского). Л., «Наука», 1968.

65. «Статистические параметры стилей» (отв. ред. В. И. Перебийнос). Киев, 1967.

66. «*Статистические и структурные лингвистические модели*». Киев, 1966 (на украинском языке).

67. Сводеш Моррис (О лексико-статистическом датировании). В кн.: «Новое в лингвистике», вып. 1. М., Изд-во иностранной литературы, 1960.

68. Сухотин Б. В. Алгоритмы лингвистической дешифровки. В кн.: «Проблемы структурной лингвистики». М., Изд-во АН СССР, 1963.

69. Творогов О. В. О применении частотных словарей в исторической лексикологии русского языка. «Вопросы языкознания», 1967, № 2.

70. Фрумкина Р. М. Статистические методы изучения лексики. М., «Наука», 1964.



71. Фрумкина Р. М. К вопросу о так называемом «законе Ципфа». «Вопросы языкознания», 1961, № 2.

72. Фрумкина Р. М. Объективные и субъективные оценки вероятностей слов. «Вопросы языкознания», 1966, № 2.

73. «Частотные словари и автоматическая переработка лингвистических текстов. Межвузовская конференция 4—6 апреля 1968 г.». Минск, 1968.

74. Чистяков В. Ф. и Крамаренко Б. М. Опыт применения статистического метода в языкознании. Вып. 1. Краснодар, 1929.

75. Шайкевич А. Я. Опыт статистического выделения функциональных стилей. «Вопросы языкознания», 1968, № 1.

76. Шеворошкин В. В. О структуре языковых цепей. В кн.: «Проблемы структурной лингвистики». М., Изд-во АН СССР, 1963.

77. Штейнфельдт Э. А. Частотный словарь современного русского литературного языка. Таллин, 1963.

78. «Энтропия языка и статистика речи» (отв. ред. Р. Г. Пиотровский). Минск, 1966.

79. «Язык и общество». Изд-во Саратовского государственного университета, 1967.

80. Арлей Н. и Бух К. Р. Введение в теорию вероятностей и математическую статистику. М., 1958.

81. Ван дер Варден Б. Л. Математическая статистика. М., 1960.

82. Венецкий И. Г., Кильдишев Г. С. Основы математической статистики. М., 1963.

83. Гнеденко В. В., Хинчин А. Я. Элементарное введение в теорию вероятностей. М., 1961.

84. Езекиэл М. и Фокс К. А. Методы анализа корреляций и регрессий. М., 1966.

85. *Йейтс Фрэнк. Выборочный* метод в переписях и обследованиях. М., 1965.

86. Митропольский А. И. Техника статистических вычислений. М., 1961.

87. «О некоторых вопросах современной математики и кибернетики. Сборник статей в помощь учителю математики». М., 1965.

88. Ряузов Н. Н. Общая теория статистики. М., 1963.

89. Смирнов Н. В., Дунин-Барковский И. В. Курс теории вероятностей и математической статистики. Для технических приложений. М., 1965.

90. Снедекор Дж. У. Статистические методы в применении к исследованиям в сельском хозяйстве и биологии. М., 1961.

91. Феллер В. Введение в теорию вероятностей и ее приложения, т. 1. М., 1967.

92. Юл Дж. Э., Кендэл М. Дж. Теория статистики. М., 1961.

93. Янко Ярослав. Математико-статистические таблицы. М., 1961.

94. Eaton H. Comparative frequency list on the first thousand words in English, French, German and Spanish. В кн.: "Experiments and studies in modern language teaching". Comp. by A. Coleman, Chicago, 1934.

95. Garcia Hoz V. Vocabulario usual, comun y fundamental. Madrid, 1953.

96. Guirand P. Problèms et métodes de la statistique linguistique. Paris, 1960.

97. Herdan G. Language as choise and chance. Groningen, 1956.

98. Herdan G. The quantitative linguistics. London, 1964.

99. Herdan G. Type — token mathematikal linguistics., S. Gravenhege, 1960.

100. T. Jakubaite, D. Kristovska, V. Ozola, R. Pruse, N. Sike. Latviešu valodas biežuma vārdnīca. 1 sēj. Riga, 1966.

101. T. Jakubaite, D. Gulevska, V. Ozola, R. Prūse, A. Rubīne, N. Sike. Latviešu valodas biežuma vārdnīca. sēj. Riga, 1969.

102. Jelinek J., Bečka J., Tešitelova M. Frekvence slov, slovnich drugů a tvarů v českém jazyce. Praha, 1961.

103. Josselson H. The Russian word count. Detroit, 1953.

104. Juilland A., Rodrigues E. Ch. Frequency dictionary of Spanish word. The Hague, 1964.

105. Juilland A., Edwards P. M., I. Juilland. Frequency distionary of Roumanian words. The Hague, 1965.

106. Kaeding F. Häufigkeitswörterbuch der Deutschen Sprache. Steigliz bei Berlin, 1898.

107. Thorndike E. a. Lorge L. The teachers word book of 30.000 words. N. Y., 1944.

108. Vacar N. P. A word count of spoken Russian. The Soviet usage. Ohio State University Press [1966].

109. Vander Beke G. French word book. N. Y., 1931.

110. Yule G. U. The statistical study of literary vocabulary. Cambridge, 1944.

## КРАТКИЙ УКАЗАТЕЛЬ ТЕРМИНОВ

Краткий указатель терминов включает те термины, которые упоминаются внутри текста. Курсивные цифры указывают те страницы, на которых даются основные определения и разъяснения данных понятий, прямые цифры указывают на главные упоминания термина.

Абсолютная ошибка наблюдения частот — *55*

Абсолютная ошибка наблюдения долей — *56*

Активность (элементов языка) — 7, 9, 54, 123—125, 142, 145

Варианты статистической методики — 3,159

Варьирование (колебание) частот — 14, *19*, *28*, 30, 32, 33

Величины $\sum a_i^2$, соответствующие числовым значениям $\chi^2$ (таблица) — 105 — 107

Величины $\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ и $\sqrt{p \cdot q}$ в зависимости от числовых значений $n_1=n_2$ и $p$ (таблица) — 112

Вероятная ошибка в определении средней частоты (ошибка наблюдения) — *25—26*, 52

Вероятностный закон — *15—16*

Вероятность — *20*, 117

Вероятность бóльшего значения t — *43—44*

Вероятность бóльшего значения $\chi^2$ — *31—32*

Вероятность подчинения имени существительному различных частей речи — 125

Вероятность применения падежей — 124

Вероятность применения частей речи — 123

Взаимосвязь вероятностей и частот — 128—129

Внутренняя зависимость между количественными и качественными характеристиками языка — 13

Выборка — 6, 62, *63 — 65*

Выборочная средняя частота — 22

Выборочная частота — 21

Выборочное значение $\chi^2$ — 30

Выборочное изучение — 51, 63—65

Выборочное отклонение — 30

Генеральная совокупность — 21
Грамматическое и лексическое описание стилевой дифференциации языка — 133—135
Границы существенности — *31*

Дисперсия — 23
Дифференциация функциональных стилей языка — 127—128
Действительная средняя частота — *45—46*
Детерминация — *166*
Дифференцирующие и нейтральные стилевые вероятности — *125—127*
Дифференцирующие и нейтральные стилевые доли — *127*
Дифференцирующие и нейтральные стилевые частоты — *130—131*
Длина (объем) выборки — 59—60
Доля — *37*

Закон Ципфа — Мандельброта — 152
Значение (числовое) $K_\phi$, деление которого на $s_{1,2}$ дает выборочное значение $t$ для сравнения двух средних частот (таблица) — 107—108

Извлечение из таблицы числовых значений $t$ — 43
Извлечение из таблицы числовых значений $\chi^2$ — 31
Измерение информационной «емкости» знаков языка — 172
Интервал действительной средней частоты — *45—46*

Квадратичное отклонение доли — *37—40*
Квадратичное отклонение разности частот — *44*

Кодовая запись текста — 66—67, 69—70
Колеблемость частот как характеристика стилей речи — 34, 132
Корреляция — *160*
Корреляционный анализ — 160
Коэффициент вариации $v$ — *35*
Коэффициент детерминации $r^2$ — *166*

Коэффициент корреляции $r$ — *161—163*

Критерий Стьюдента — *41*
Критерий согласия «хи-квадрат» ($\chi^2$) — *28—29*

Критерий порядковый «хи» (X) — *47*

Лингвоструктура — *129*
Лингвистическая интерпретация выборочных частот и долей — 88—90

Минимальные величины истинных коэффициентов корреляции (таблица) — 164
Минимальные значения $\sum a_{ti} + \sum a_{t_2}^2$, соответствующие несущественной разности двух средних частот (таблица) — 107—108

Надежность определения ошибки наблюдения — 53
Надежность определения средней частоты — *60—61*
Несмещенная оценка среднего квадратичного отклонения — 25, 26, 42
Нулевая гипотеза — *47*

Общие вопросы статистического изучения языка — 170—171
Объективная необходимость статистического изучения языка — *16—17*
Объективность количественных характеристик языка — *11*
Отклонение выборочных частот от их средней частоты — 23
среднее абсолютное отклонение — *23*
среднее квадратичное отклонение — *23*
Отклонение (расхождение) частот — 28, 29, 30, 35
Отклонение (расхождение) долей — 37
Относительная ошибка наблюдения долей — *56—57*
Относительная ошибка наблюдения частот — *55*
Определение (планирование) достаточного числа выборок (наблюдений) — 57—58
Определение (планирование) достаточного объема (длины) выборок — 59—60

Планирование статистического эксперимента — 57, 62
Применение критерия «хи-квадрат» ($\chi^2$) — *32—33*
Программы статистического изучения языка и речи — 66
Программа «Части русского языка» — 67—68
Программа «Имя существительное» — 72—73
Программа «Имя прилагательное» — 74—76
Программа — «Глагол» — 76—78
Программа «Простое предложение» — 78—79
Программа «Сложное предложение» — 80—81

Ранжированный ряд (частот) — 47
Речевая вероятность — 115—117

Случайное расхождение частот — 30
Соотношение между надежностью и точностью — 61
Союз статистики с качественными (классическими) методиками изучения языка и речи — *17*
Сравнение долей — *36—37*
Сравнение средних выборочных частот — *40*
Сравнение частотных рядов — *46—47*
Средняя выборочная частота — *22*
Статистика в атрибуции литературных текстов — 155 — 156
Статистика в дешифровке древних текстов — 172
Статистика в изучении истории языка — 157—159
Статистика в изучении речевой культуры — 143
Статистика в изучении речи школьника — 154—155
Статистика в изучении стиха — 171
Статистика в изучении художественной речи — 136—143
Статистический (вероятностный) закон — 14—17, 19
Статистическое исследование языка — 52, 62—63
Статистическая лингвистика — 172
Статистическое описание языка — 51, 62—63
Статистическая оценка расхождений между выборочными частотами (сравнение частот) — 27, *39—40*

Статистическое равенство частот (и долей) — 30, 38
Статистическая стратификация текста — 90—91
Статистическая таблица, фиксирующая результаты статистических наблюдений — 83—87
Стилевая дифференциация языка — *133*
Стилостатистика — 169—170
Стиль автора — 5, 8, 128, 130—131, 139—143, 149—150
Стиль речи—*129—132*
Стиль языка — *118—122*
Существенно-различное соотношение частот — *132*
Существенное расхождение частот — *30*

Таблицы, облегчающие статистические вычисления — 91—92
Тенденции функционального притяжения и отталкивания частей речи — 165—166
Точность определения средней частоты (доли) — 60—61

Условия применения статистики в изучении языка и речи — 17—19
Формальные качества речи — мысли — 143
  богатство — 148, 151—153
  динамизм — 147
  качественность — 147
  предметность — 147
  разрывность — 147—148
  расчлененность — 146
  ровность — 147
  связанность — 145
  сложность — 148—151
  темп — 144—145
  уточненность — 145—146

Частота — 21
Частотный словарь — 167—169
Число степеней свободы — 29—31
Числовые значения $L$ и $s$ в зависимости от $\sum a_i^2$ (таблица) — 93—105
Числовые значения $\psi$ (таблица) — 48

Язык и речь — 128—129, 133

## СОДЕРЖАНИЕ

От автора . . . . . 3
Вместо введения . . . . . 5
Основания и условия вероятностно-статистического изучения языка и речи . . . . . 10
Минимально-необходимые статистические инструменты . . . 19
Статистическая оценка расхождений между выборочными частотами . . . . . 28
Сравнение долей . . . . . 36
Сравнение средних выборочных частот и частотных рядов . 40
Ошибки наблюдения и определение объема выборок из текста 50
Организация статистического изучения языка и речи . . . . 63
Таблицы, облегчающие статистические вычисления . . . . 91
Учение о стилях языка и стилях речи и статистика . . . . . 113
Проблемы и перспективы . . . . . 133
Послесловие . . . . . 167
Приложение 1. Квадраты целых и дробных чисел от единицы до десяти . . . . . 174
Приложение 2. Квадратные корни из целых и дробных чисел от единицы до ста . . . . . 177
*Основная литература . . . . . 182*
Краткий указатель терминов . . . . . 187

*БОРИС НИКОЛАЕВИЧ ГОЛОВИН*

**ЯЗЫК И СТАТИСТИКА**

*Редактор Г. В. Кармаз*
Художественный редактор *Н. А. Володина*
Технический редактор *М. Н. Смирнова*
Корректор *Т. М. Графовская*

* * *

Сдано в набор 27/V 1970 г. Подписано к печати 20/XI 1970 г. 84×108¹/₃₂.
Типографская № 3.
Печ. л. 6. Услови. л. 10,08 Уч.-изд. л. 10,48.
Тираж 25 тыс. экз. (Тем. пл. 1971 г. БЗ №. 15—1971—№ 14). А038 18.

Издательство «Просвещение» Комитета по печати при Совете Министров РСФСР. Москва, 3-й проезд Марьиной рощи, 41.

Типография издательства «Уральский рабочий», Свердловск, пр. Ленина, 49.
Заказ № 359

**Цена 29 коп.**