ДНЕПРОПЕТРОВСКИЙ УНИВЕРСИТЕТ
имени АЛЬФРЕДА НОБЕЛЯ

Ю.К. Тараненко
О.Б. Тарнопольский

ЛИНГВИСТИЧЕСКАЯ СТАТИСТИКА

ДНЕПРОПЕТРОВСКИЙ УНИВЕРСИТЕТ
имени АЛЬФРЕДА НОБЕЛЯ

Ю.К. ТАРАНЕНКО
О.Б. ТАРНОПОЛЬСКИЙ

ЛИНГВИСТИЧЕСКАЯ СТАТИСТИКА

УЧЕБНОЕ ПОСОБИЕ

Днепропетровск
2014

УДК 81’33
ББК 81.1я7
Т 19

Рецензент:
В.М. Косарев, кандидат технических наук,
профессор кафедры экономической кибернетики и математических методов в экономике Днепропетровского университета имени Альфреда Нобеля

У навчальному посібнику вперше на підставі методів статистичного дослідження мови і мовлення, розроблених Б.А. Головіним, наведено автоматизацію зазначених методів із застосуванням орієнтованої на розв’язання лінгвістичних задач потужної мови програмування Python 3.40 з набором інструментів природної мови NLTK. Цей навчальний посібник орієнтовано на студентів спеціальності «Прикладна лінгвістика». Практичним виданням до посібника є збірник задач з лінгвістичної статистики, який містить лістинги авторських програм на Python 3.40. Разом із збірником задач посібник являє собою методичний комплекс, який може бути корисним у роботі прикладного лінгвіста, а також аспірантів і викладачів.

Тараненко Ю.К.

Т 19 Лингвистическая статистика: учебное пособие / Ю.К. Тараненко, О.Б. Тарнопольский. – Днепропетровск: Днепропетровский университет имени Альфреда Нобеля, 2014. – 112 с.

ISBN 978-966-434-325-8

В учебном пособии впервые на основе методов статистического исследования языка и речи, разработанных Б.А. Головиным, приведена автоматизация указанных методов, с применением ориентированного на решение лингвистических задач мощного языка программирования Python 3.40 с набором инструментов естественного языка NLTK. Данное учебное пособие ориентировано на студентов специальности «Прикладная лингвистика». Практическим изданием к пособию является сборник задач по лингвистической статистике, который содержит листинги авторских программ на Python 3.40. Вместе с задачником пособие представляет собой методический комплекс, который может быть полезен в работе прикладного лингвиста, а также аспирантов и преподавателей.

УДК 81’33
ББК 81.1я7

ISBN 978-966-434-325-8

СОДЕРЖАНИЕ

ТЕМА 1. Введение в лингвистическую статистику 4

ТЕМА 2. Минимально необходимые статистические инструменты 11

ТЕМА 3. Статистическая оценка расхождений между выборочными частотами 25

ТЕМА 4. Сравнение долей 39

ТЕМА 5. Сравнение средних выборочных частот и частотных рядов 47

ТЕМА 6. Ошибки наблюдения и определение объема выборок из текста 66

ТЕМА 7. Организация статистического изучения языка и речи на основе современных информационных технологий 80

Приложение № 1. Функции класса FreqDist 108

Приложение № 2. Перечень корпусов текстов, которые распространяются вместе с NLTK 110

Литература 112

ТЕМА 1. ВВЕДЕНИЕ В ЛИНГВИСТИЧЕСКУЮ СТАТИСТИКУ

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 стиле автора. Часто по двум-трем предложениям читатель устанавливает, кем они могли быть написаны.

Можно брать совершенно случайно, не обращая внимания на содержание, отрывки 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 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 независимо от быстротекущего и изменчивого содержания.

А если это так, разве может лингвист (да и литературовед) не заинтересоваться теми возможностями, которые открывает систе-

матическое изучение языкового функционирования и развития при помощи статистики? Это ставит ряд весьма сложных вопросов, требующих творческого решения. Среди них:

а) связаны ли показываемые статистикой особенности функционирования частей речи, предложений и их членов в речи, например, Симонова и Шолохова некоторыми внутренними зависимостями, т. е. **носят ли они системный характер?**

б) стоят ли за различиями активности частей речи, членов предложения 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, явлении лексических и лексико-семантических?

г) следует ли думать, что и в неисследованных 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 вв.?

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 – накопления фактов, 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, требуются и отказ от предвзятости, и внутренняя решимость, и усилия, и время. Правда, все эти «потери» с лихвой окупятся получаемыми с помощью статистики ре-

зультатами – 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

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.

Вторым основанием является **внутренняя зависимость, существующая между качественными и количественными характеристиками языковой структуры** [1].

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 толь-

ко 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».

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, 42У 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, 68У 80, 95, 112, 68, 76У 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 зако-

номерностях, 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 колебании 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 и речь образуются 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].

Выводы

1. Союз статистики с традиционными методиками качественного анализа языка 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.

2. Вторым условием успешного применения статистики в науке о языке представляется более или менее отчетливое понимание ученым типов лингвистических 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.

3. Третье условие успеха в применении статистической методики – знакомство филолога с минимально необходимыми для этого статистическими инструментами.

Контрольные вопросы

1. Что в структуре речи не зависит от быстротекущего и изменчивого содержания?

2. Почему читатель узнаёт авторство даже по отдельным отрывкам произведения?

3. Какие возможности открывает систематическое изучение языкового функционирования и развития при помощи статистики?

4. Что является реальным основанием применения статистики для изучения языка и речи?

5. Как соотносятся качественные и количественные характеристики языковой структуры?

6. Каким законам подчиняются частоты различных элементов языка в речевом потоке?

7. Частоты каких частей речи можно использовать для статистического анализа творчества писателя?

8. Поясните, какие законы природы являются динамическими.

9. Поясните, какие законы природы являются статистическими.

10. Какие основные выводы можно сделать из материала темы?

Практическое задание № 1

1. Выбрать в сети Internet не менее 5 фрагментов текста по 100 словоупотреблений для одного автора.

2. Сделать предварительный вывод об общих элементах.

3. Сохранить результаты на флэш-карте.

ТЕМА 2. 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 на первой лекции. 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 воз-

действий, 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 изменения в сочетании многих воздействий, то величина этих отклонений подчинена определенному закону, который и может быть установлен, и описан с помощью математики. 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): $P = m / n$.

Когда мы подбрасываем много раз игральный кубик, мы можем заранее, до исхода нашего опыта сформулировать гипотезу о равной вероятности выпадения 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 уста-

новить, сохраняется или нарушается эта вероятность, 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, 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 сторона с отметкой «один» 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

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^2}{k}};$$

где σ – среднее квадратичное отклонение, $(x_i - \bar{x})$ – отклонения выборочной частоты от средней; $\sum_{i=1}^{k}$ – знак суммирования этих отклонений; k – число выборок (наблюдений); если примем, $x_i - \bar{x} = a_i$ то формулу можно записать в более простом виде:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} a_i^2}{k}}.$$

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).**

Формулу $\sigma = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2}{k}}$, можно также упростить следующим образом:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} x_i^2}{k} - \bar{x}^2\left(\frac{2\cdot\left(\sum_{l=1}^{k} x_L\right)^2}{\bar{x}^2\cdot k^2} - 1\right)} = \sqrt{\frac{\sum_{i=1}^{k} x_i^2}{k} - \bar{x}^2}, \quad \sigma^2 = \frac{\sum_{i=1}^{k} x_i^2}{k} - \bar{x}^2.$$

Кстати, два попутных замечания: 1) формула сообщена здесь в своем простейшем виде для случая, когда все выборки равны по

длине или объему; 2) величина $\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2}{k}$ носит в статистике название дисперсии и обозначается знаком σ^2.

В математической статистике пользуются обычно не средним абсолютным, а именно средним квадратичным отклонением – из-за его чисто математических преимуществ, рассматривать которые здесь не место.

Пример 2.1. 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 33,2; 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. Поскольку из всех языков программирования наиболее приемлем для аналитики текстов Python. Запишем решение примера 1 в виде скрипта sko.py.

Текст скрипта sko.py

```
a=[95,87,94,104,100]
s = 0
for x in a:
        s =s+x # накопления суммы элементов массива частот
print('Сумма элементов массива частот- '+str(s))
```

```
r=s/len(a) # len(a) -число элементов в массиве частота
print('Средняя частота - '+str(r))
s = 0
for x in a: # накопление суммы квадратов отклонения текущих частот от средней
        s =s+ (x - r) ** 2
d= round(s/(len(a)),2)
print('Дисперсия частот - '+str(d))
sko=round(d**0.5,2)
print('Среднее квадратичное отклонение - '+str(sko))
```

Результат работы скрипта

Сумма элементов массива частот – 480.0
Средняя частота – 96.0
Дисперсия частот – 33.2
Среднее квадратичное отклонение – 5.76.

Математическая статистика утверждает, что в практике статистического изучения лучше применять не формулу $\sigma = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2}{k}}$, а другую, отличающуюся только тем, что в знаменателе дроби под знаком корня стоит не k, а $k - 1$ – $\sigma = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2}{k-1}}$. 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 нужно было 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 (о том, как это делается, 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:

$$L = \frac{t \cdot \sigma}{\sqrt{k}},$$

где L – величина ошибки, t – особый коэффициент, зависящий от числа наблюдений (выборок), он берется из таблицы; σ – знакомое нам среднее квадратичное отклонение или вместо него мож-

но взять $s = \sqrt{\frac{\sum_{i=1}^{k} a_i^2}{k-1}}$ (это даже лучше); k – число наблюдений (выборок).

При пяти выборках t 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 А 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 А 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 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.

Пример 2.2. 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! 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 $L = \frac{2 \cdot s}{\sqrt{k}}$ (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 ТВ – от 101 до 119 (результаты вычислений округлены). Запишем решение примера 2 в виде скрипта L.py.

Текст скрипта L.py

```
a=[95, 98, 89, 105, 102, 85, 111, 115, 93, 107]
s = 0
for x in a:
        s =s+x # накопления суммы элементов массива частот
print('Сумма элементов массива частот выборки1 - '+str(s))
r=round(s/len(a),0) # len(a) -число элементов, round(,0) округление
до целых
print('Средняя частота выборки1 – '+str(r))
s = 0
for x in a:
        s =s+ (x – r) ** 2 # накопление суммы квадратов отклонения
        текущих частот от средней
d= round(s/(len(a)-1),0)
```

```
print('Дисперсия частот выборки 1- '+str(d))
sko=round(d**0.5,0)
print('Среднее квадратичное отклонение выборки 1 - '+str(sko))
l=round(2*sko/(len(a))**0.5,0)
print('Ошибка в определение средней частоты выборки 1 - '+str(l))
b=[98, 112, 114, 108, 106, 122, 95, 87, 125, 133]
s1= 0
for x in b:
        s1 =s1+x # накопления суммы элементов массива частот
print('Сумма элементов массива частот выборки 2 - '+str(s1))
r1=round(s1/len(b),0) # len(a) -число элементов в массиве частот
print('Средняя частота выборки 2 - '+str(r1))
s1= 0
for x in b: # накопление суммы квадратов отклонения текущих ча-
стот от средней
        s1 =s1+ (x - r1) ** 2
d1= round(s1/(len(b)-1),0)
print('Дисперсия частот выборки 2 - '+str(d1))
sko1=round(d1**0.5,0)
print('Среднее квадратичное отклонение выборки 2 - '+str(sko1))
l1=round(2*sko1/(len(b))**0.5,0)
print('Ошибка в определения средней частоты выборки 2- '+str(l1))
print('Диапазоны в которых находиться средняя частота\n\
для выборк 1,2: '+str(r-l)+'--'+str(r+l)+' , '+str(r1-l1)+'--'+str(r1+l1))
```

Результат работы скрипта

```
Сумма элементов массива частот выборки 1 – 1000.
Средняя частота выборки 1 – 100.0.
Дисперсия частот выборки 1 – 94.0.
Среднее квадратичное отклонение выборки 1 – 10.0.
Ошибка в определения средней частоты выборки 1 – 6.0.
Сумма элементов массива частот выборки 2 – 1100.
Средняя частота выборки 2 – 110.0.
Дисперсия частот выборки 2 – 204.0.
Среднее квадратичное отклонение выборки 2 – 14.0.
Ошибка в определении средней частоты выборки 2 – 9.0.
Диапазоны, в которых находиться средняя частота
для выборки 1,2: 94.0 –106.0; 101.0 – 119.0.
```

Контрольные вопросы

1. Сколько раз при 600 подбрасываниях выпадет одна сторона шестигранного кубика?

2. Какое средство сокращения научного эксперимента позволяет по нескольким пробам, выборкам судить о совокупности явлений в целом?

3. Дайте общее определение вероятности события.

4. Запишите соотношение для вероятности события.

5. Поясните связь между статистическим (вероятностным) законом и вероятностью.

6. Что нужно отнести к числу самых элементарных инструментов наблюдения за действием статистических законов, за вероятностью?

7. Приведите примеры частот при анализе литературных произведений.

8. Чем определяется однородность выборок частей текста. Приведите примеры.

9. Какими путями достигается обобщающее усреднение в статистике

10. Приведите все три формы записи для формулы среднего квадратичного отклонения.

11. Приведите формулу для несмещенной оценки среднего квадратичного отклонения.

12. Что характеризует дисперсия для выборки частот.

13. Приведите вывод формулы $\sigma^2 = \frac{\sum_{i=1}^{k} x_i^2}{k} - \bar{x}^2$ для дисперсии.

14. Опишите словестно формулу для определения среднего квадратичного отклонения.

15. Что определяется термином «вероятная ошибка» в определении средней

частоты.

16. Приведите формулу для определения вероятной ошибки средней частоты.

17. Какая надёжность необходима для определения ошибки средней частоты.

18. Поясните работу скриптов sko.py,L.py.

Практическое задание № 2

(Пример общей методики исследования скриптов Python)

1. Запустите IDLE (Python GUI). Откроется окно:

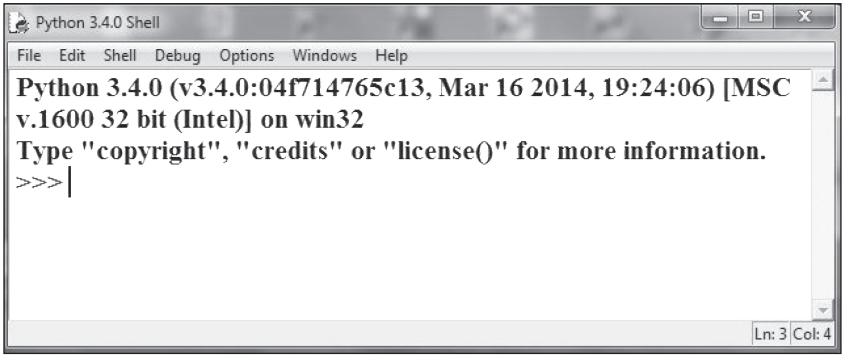

Рис. 2.1. Основное окно графического интерфейса Python 3.4

2. В меню File выберите NewFile или нажмите Ctr+N. Откроется окно, в которое скопируйте текст скрипт asko.py

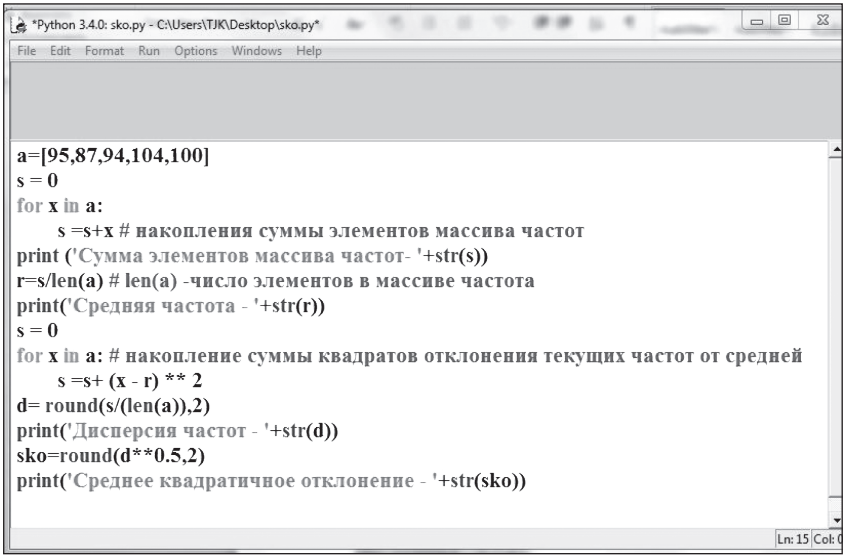

Рис. 2.2. Окно редактора скрипта

3. Выберите в меню Run подменю RunModule или нажмите F5.

Сохраните скрипт в файле под именем sko с автоматически устанавливаемым расширением – ру. Убедитесь в том, что результат соответствует примеру 1.

4. Вернитесь в текст скрипта и измените половину чисел в списке: a=[95, 98, 89, 105, 102, 85, 111, 115, 93, 107]. Запустите скрипт.

5. Скопируйте результат в отчёт. Отчёт составьте как в примере 1 из двух частей: 1. Текст скрипта _ .ру: 2. Результат работы скрипта: Добавьте третьим пунктом «Выводы», в которых поясните, какие параметры и почему изменились.

6. Аналогичные действия проделайте для скрипта L.ру. Результаты предъявите преподавателю и сохраните на флэш-карте.

ТЕМА 3. 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 [1]. 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, 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 χ^2. Вот формула, по которой вычисляется величина «хи-квадрат» в таких случаях, как наш, т. е. когда все выборки имеют одинаковую длину:

$$\chi^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\bar{x}},$$

где x_i – наблюдаемые частоты; $\bar{x}$ – средняя выборочная частота; $\sum_{i=1}^{n}$ – знак суммирования.

Если обозначить отклонение выборочной частоты от средней буквой «a_i», как это было сделано ранее, то формула «хи-квадрат» получит менее громоздкий вид $\chi^2 = \frac{\sum_{i=1}^{n} (a_i)^2}{\bar{x}}$. 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, соответствующим 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 которой 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) определенной вероятности (вертикальный столбец). Понятие «степень свободы» оставим пока без пояснений ввиду его сложности; но примем без дока-

зательств, что при сравнении нескольких выборочных частот – и при условии, что все выборки имеют равную длину, – число степени 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 вероятности – 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 большую, встречается в пяти теоретических случаев 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, то 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?

Пример 3.1. Допустим, что были сделаны 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. Запишем решение примера 2 в виде скрипта psi.py.

Текст скрипта psi.py

```
a=[75, 70, 82, 68 , 80]
s = 0
for x in a:
        s =s+x # накопления суммы элементов массива частот
print('Сумма элементов массива частот выборки - '+str(s))
r=round(s/len(a),2) # len(a) -число элементов в массиве частот
print('Средняя частота выборки - '+str(r))
s = 0
for x in a: # накопление суммы квадратов отклонения текущих частот от средней
        s =s+ (x – r) ** 2
d= round(s/(len(a)-1),2)
print('Дисперсия частот выборки - '+str(d))
psi=round(s/r,2)
print('Расчётный ПСИ-критерий частот выборки - '+str(psi))
```

Результат работы скрипта

```
Сумма элементов массива частот выборки – 375.
Средняя частота выборки – 75.0.
Дисперсия частот выборки – 37.0.
Расчётный ПСИ-критерий частот выборки – 1.97.
```

В табл. 3.1, выбираем строку, соответствующую четырем степеням свободы только четыре частоты из пяти могут в формуле «хи-квадрат» принимать любые значения, если дана еще и сред-

няя 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.

Таблица 3.1

Числовые значения χ^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,П	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

Почему? Потому что 95-процентная вероятность слишком велика: соответствующая ей или меньшая величина «хи-квадрат» встречается всего пять раз на сто, и 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. 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, что 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 χ^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. Всё 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 колебаний 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 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 возник-

ла. Величиной колеблемости (а она может оцениваться величиной «хи-квадрат») характеризуется устойчивость или неустойчивость различных элементов языковой структуры в разных условиях их текстового применения.

Пример 3.2. В опыте получены две частоты одного и того же явления языка в двух текстовых совокупностях, выборки из которых были равного объема (выборки, разумеется, могут «отмерять-ся» 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. суммируем наши частоты и разделим пополам. Пусть, например, частоты были 270 и 220, их средняя – 245. Затем применим формулу «хи-квадрат»:

$$\chi^2 = \frac{(270 - 245)^2}{245} + \frac{(220 - 245)^2}{245} = 5,1.$$

Так как была использована одна степень свободы, таблица покажет нам, что критическая величина «хи-квадрат» равна 3,84; полученная из опыта превосходит критическую, причем значительно. Вероятность такого значения величины «хи-квадрат», которое нам дал опыт, весьма невелика, заметно менее 5%. Поэтому гипотезу о несущественности, случайности расхождения частот приходится отвергнуть.

Для автоматизации решения задач о характере отклонения средних частот (СЛУЧАЙНЫЙ, НЕ СЛУЧАЙНЫЙ), при двух выборках равного объёма приведём скрипт psi12.py.

Текст скрипта psi12.py

```
a=float(input('Введитесреднююючастоту первой выборки: '))
b=float(input('Введите среднюю частоту второй выборки: '))
k=(a+b)/2
psi=((a-k)**2/k)+((b-k)**2/k)
c=3.85
if psi<c:
```

```
        print('С вероятностью 0,05 расчётная величина пси-'
        +str(round(psi,2))+'\n\
меньше теоритической табличной '+str(c)+'\n\
Отклонение частот от средней -' +str(round(k,2))+'- СЛУЧАЙНОЕ ')
if psi>=c:
        print('С вероятностью 0,05 расчётная величина пси-'
+str(round(psi,2))+'\n\
больше теоритической табличной '+str(c)+'\n\
Отклонение частот от средней -' +str(round(k,2))+'- НЕ СЛУЧАЙНОЕ ')
```

Результат работы скрипта

```
Введите среднюю частоту первой выборки: 270.
Введите среднюю частоту второй выборки: 220.
С вероятностью 0,05 расчётная величина пси-5.1
больше теоритической табличной 3.85.
Отклонение частот от средней – 245.0 – НЕ СЛУЧАЙНОЕ.
```

Пример 3.3. Несколько меняется та же по существу задача, если выборки из текста не были строго одинаковыми, например, одна равнялась 530 знаменательным словам, а другая – 970. Как поступить в этом случае? Очевидно, что для получения выборочной средней частоты уже нельзя сложить показанные двумя выборками частоты и разделить их пополам. Пусть эти частоты, скажем, имен прилагательных равнялись 75 и 100. Суммируем эти частоты, суммируем обе выборки, разделим сумму частот на величину совокупной выборки и затем умножим результат деления на величину 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 χ^2, поданным 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.

Для автоматизации решения задач о характере отклонения средних частот (СЛУЧАЙНЫЙ, НЕ СЛУЧАЙНЫЙ), при двух выборках **разного объёма**, приведём скрипт psi12r.py.

Текст скрипта psi12r.py

```
a=float(input('Введите среднюю частоту первой выборки: '))
b=float(input('Введите среднюю частоту второй выборки: '))
e=float(input('Введите объём первой выборки: '))
f=float(input('Введите объём второй выборки: '))
k1=((a+b)/(e+f))*e
k2=((a+b)/(e+f))*f
psi=((a-k1)**2/k1)+((b-k2)**2/k2)
c=3.85
if psi<c:
        print('Свероятностью 0,05 расчётная величина пси-'
+str(round(psi,2))+'\n\
меньше теоретической табличной '+str(c)+'.\n\
Отклонение средних частот – СЛУЧАЙНОЕ ')
if psi>=c:
        print('Свероятностью 0,05 расчётная величина пси-'
+str(round(psi,2))+' \n\
больше теоретической табличной '+str(c)+'.\n\
Отклонение средних частот – НЕ СЛУЧАЙНОЕ ')')
```

Результат работы скрипта

```
Введите среднюю частоту первой выборки: 75.
Введите среднюю частоту второй выборки: 100.
Введите объём первой выборки: 530.
Введите объём второй выборки: 970.
С вероятностью 0,05 расчётная величина пси-4.34
больше теоретической табличной 3.85.
Отклонение средних частот – НЕ СЛУЧАЙНОЕ.
```

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 сходной задачей сравнения долей, что иногда бывает даже удобнее, как, например, в случаях, когда выборки заметно отличаются по величине друг от друга.

Посмотрим, что собой представляет и как применяется коэффициент вариации. Для этого припомним понятие и термин «среднее квадратичное отклонение», его формула, как уже говорилось, такова:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2}{k}},$$

Зная среднее квадратичное отклонение, мы сравнительно легко найдем коэффициент вариации по формуле:

$$v = \frac{100 \cdot \sigma}{\bar{x}}.$$

Нетрудно увидеть, что коэффициент вариации представляет собой не что иное, как отношение среднего квадратичного отклонения к средней частоте, выраженное в процентах. Это отношение, естественно, тем больше, чем больше среднее квадратичное отклонение и чем меньше средняя частота. Но квадратичное отклонение обобщающе-усредненно характеризует совокупность отклонений выборочных частот от их средней частоты, поэтому и коэффициент вариации, показывая отношение σ к $\bar{x}$, является удовлетворительной мерой колеблемости частот, применяемой статистиками при решении ряда практических задач.

Пример 3.4. 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 существенно-

сти, 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 автоматизации решения задач о характере частот в выборке вокруг средней частоты (СЛУЧАЙНЫЙ, НЕ СЛУЧАЙНЫЙ) с использованием коэффициента вариации приведём скрипт var.py.

Текст скрипта var.py

```
a=[98, 87, 102, 105, 123, 108, 85, 78, 110, 104]
s = 0
for x in a:
        s =s+x # накопления суммы элементов массива частот
print ('Сумма элементов массива частот- '+str(s))
r=s/len(a) # len(a) -число элементов в массиве частота
print('Средняя частота - '+str(r))
s = 0
for x in a: # накопление суммы квадратов отклонения текущих частот от средней
        s =s+ (x – r) ** 2
d= round(s/(len(a)),2)
print('Дисперсия частот – '+str(d))
sko=round(d**0.5,2)
print('Среднее квадратичное отклонение – '+str(sko))
var=(sko/r)*100
if var<40:
        print('Коэффициент вариации '+str(round(var,2))+' % – меньше 40%.\n\
Разброс частот около среднего носит СЛУЧАЙНЫЙ характер')
if var>=40:
        print('Коэффициент вариации '+str(round(var,2))+' % – больше (равен) 40%.\n\
Разброс частот около среднего носит НЕ СЛУЧАЙНЫЙ характер')
```

Результат работы скрипта

Сумма элементов массива частот – 1000.
Средняя частота – 100.0.
Дисперсия частот – 162.0.
Среднее квадратичное отклонение – 12.73.
Коэффициент вариации 12.73% меньше 40%.
Разброс частот около среднего носит СЛУЧАЙНЫЙ характер.

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.

Контрольные вопросы

1. Как ведётся статистическое изучение языка или речи?
2. Что можно сказать о выборке частот словоупотреблений из текста, когда колебания частот в выборке из текста около среднего носят случайный характер?
3. Что можно сказать о выборке частот словоупотреблений из текста, когда колебания частот в выборке из текста около среднего носят не случайный характер?
4. Как используется «хи-квадрат критерий» при статистическом изучении текста?
5. Запишите формулу, по которой вычисляется величина «хи-квадрат», когда все выборки имеют одинаковую длину.
6. Опишите формулу для вычисления величины «хи-квадрат» словесно.
7. Критерий «хи-квадрат» часто называют критерием согласия. Чего с чем?

8. Как определяют степень свободы при выборе табличного значения величины «хи-квадрат»?

9. Поясните понятие «вероятность большего значения» величины «хи-квадрат» с применением таблицы числовых значений χ^2.

10. Поясните понятие «границы существенности» с применением таблицы числовых значений χ^2.

11. Сколько раз из 100 встречаются меньшие значения χ^2 при 95% вероятности больших значений?

12. Какой вывод следует из того что в первом же опыте мы получили расчётную величину χ^2 меньше табличного при 5% вероятности больших значений.

13. Какой вывод следует из того, что в первом же опыте мы получили расчётную величину χ^2 больше табличного при 95% вероятности больших значений?

14. Когда принимают гипотезу о случайности колебания наблюдавшихся частот около их общей средней частоты?

15. Какой критерий даёт вполне определенный ответ на вопрос «Верна или не верна данная статистическая гипотеза?».

16. Применение какого критерия наиболее удобно при определении величины и характера колебания частот около средней?

17. К чему может привести то, что «хи-квадрат» обладает такой большой точностью, такой большой бракующей гипотезы силой?

18. Какой критерий можно рекомендовать лингвистам для оценки колебания частот около средней, когда сравниваемые частоты находятся в промежутке от десятка-двух до нескольких сотен?

19. Поясните работу скрипта psi.py.

20. Поясните работу скрипта psi12.py.

21. Поясните работу скрипта psi12r.py.

22. Напишите формулу для коэффициента вариации.

23. Поясните работу скрипта var.py.

Практическое занятие № 3

Задание: Исследовать скрипты psi.py, psi12.py, psi12r.py, var.py по общей методике исследования скриптов Python (Практическое задание № 2. Пример). При изменении исходных данных там, где программа определяет характер рассеивания частот возле средней частоты подбором исходных данных добивайтесь противоположного по отношению к шаблону результата. Для каждого такого случая дайте детальное пояснение.

ТЕМА 4. 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 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:

$$\varepsilon_{1,2} = \sqrt{\overline{p} \cdot \overline{q} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)},$$

где $\varepsilon_{1,2}$ – величина квадратичного отклонения средней доли двух сравниваемых совокупностей; $\overline{p}$ и $\overline{q}$ – средние (для двух совокупностей) доли изучаемых явлений и всех остальных; n_1 и n_2 – размеры выборок.

Пример 4.1. 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 p и q. Так как выборки были равны, то p мы получим, сложив 0,20 и 0,15 (доли глаголов в двух выборках) и разделив сумму пополам (получится $\overline{p} = 0{,}175$; следовательно, $\overline{q} = 0{,}825$). 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 – 0,05 = 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 $3 \cdot \varepsilon_{1,2} < p_1 - p_2$, то расхождение долей существенно. В нашей задаче квадратичная ошибка средней доли 0,017, а утроенная 0,051 больше разности долей 0,05 всего на 0,001. Поэтому до повторного, уточняющего, опыта нельзя принять гипотезу о несущественности, случайности того расхождения долей, которое показали выборки. Для автоматизации решения задач о характере колебания долей в выборке вокруг средней доли (СЛУЧАЙНЫЙ, НЕ СЛУЧАЙНЫЙ) с использованием утроенной величины квадратичного отклонения средней доли двух сравниваемых совокупностей приведём скрипт dol.py

Текст скрипта dol.py

```
a=float(input('Введите объём выборок долей: '))
b=float(input('Введите количество словоформ в первой выборке: '))
c=float(input('Введите количество словоформ во второй выборке: '))
p=round((b+c)/(2*a),2)
q=1-p
e=round((p*q*2/a)**0.5,3)
k=round(abs((b/a)-(c/a)),2)
m=round(3*e,3)
if m>k:
        print('Утроенное квадратичное отклонение долей -'+str(m)+ '
больше их разности -'+str(k)+'\n\
рассеивание долей около среднего – '+str(p)+' носит случайный ха-
рактер.')
if 3*e<=k:
        print('Утроенное квадратичное отклонение долей -'+str(m)+
'меньше их разности -'+str(k)+'\n\
рассеивание долей около среднего -'+str(p)+' носит неслучайный
характер.')
```

Результат работы скрипта

```
Введите объём выборок долей: 1000.
Введите количество словоформ в первой выборке: 200.
Введите количество словоформ во второй выборке: 150.
Утроенное квадратичное отклонение долей -0.051 больше их раз-
ности –0.05,
рассеивание долей около среднего – 0.17 носит случайный характер.
```

Ван дер Варден [2; стр. 275–276] рекомендует для сравнения долей (вероятностей) применять критерий «хи-квадрат». Одна из формул, пригодных к действию, такова:

$$\chi^2 = \frac{(x_1 - n_1 \cdot \overline{p})^2}{n_1 \cdot \overline{p} \cdot \overline{q}} + \frac{(x_2 - n_2 \cdot \overline{p})^2}{n_2 \cdot \overline{p} \cdot \overline{q}},$$

где x_1 и x_2 – выборочные частоты; $\overline{p}$ – средняя доля частот; $\overline{q}$ – 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:

$$\chi^2 = \frac{(x_1 - \bar{x})^2}{\bar{x} \cdot \bar{q}} + \frac{(x_2 - \bar{x})^2}{\bar{x} \cdot \bar{q}},$$

где x_1 и x_2 – выборочные частоты, $\bar{p}$ – их средняя, подучаемая делением их суммы пополам, q – выборочная средняя доля всех единиц в выборках, кроме наблюдаемых, т. е. кроме обозначенных 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. 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 q не зависит от x.

Пример 4.3. Проведём опыт параллельного решения статистических задач с помощью формулы ошибки средней доли и критерия «хи-квадрат». В формуле «хи-квадрат» 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, **Пример 4.1**) и то же расхождение **существенно** (говорит формула «хи-квадрат» при допуске одной степени свободы). Воспользуемся скриптом **dol.py**.

Результат работы скрипта dol.py

Введите объём выборок долей: 1000.
Введите количество словоформ в первой выборке: 200.
Введите количество словоформ во второй выборке: 160.
Утроенное квадратичное отклонение долей – 0.051 больше их разности – 0.04,
рассеивание долей около среднего – 0.18 носит **случайный характер.**

Для анализа применения критерия «хи-квадрат» создадим скрипт **dol_hi2.py.**

Текст скрипта dol_hi2.py

```
a=float(input('Введите объём выборок долей: '))
b=float(input('Введите количество словоформ в первой выборке: '))
c=float(input('Введите количество словоформ во второй выборке: '))
n=int(input('Введите число степеней свободы: '))
p=round((b+c)/(2*a),4)
q=1-p
psi=round(((b-a*p)**2/(a*p*q))+((c-a*p)**2)/(a*p*q),3)
t=[3.84,5.95,7.81,9.49,11.07,12.59,14.07,15.51,16.92,18.31,23.68,25.0
0,30.14,31.41,36.42,37.65,42.56,43.77]
k=n-1
t1=t[k]
if psi<t1:
        print('С вероятностью 0,05 величина критерия будет боль-
ше табличного – \n\
расчётная величина пси-' +str(psi)+'меньше теоретической таблич-
ной '+str(t1)+'.\n\
Отклонение долей от средней доли -' +str(p)+'- носит случайный ха-
рактер')
if psi>=t1:
        print('С вероятностью 0,05 величина критерия будет боль-
ше табличного – \n\
расчётная величина пси-' +str(psi)+' больше или равна теоретиче-
ской табличной '+str(t1)+'\n\
Отклонение долей от средней доли -' +str(p)+' носит неслучайный
характер ')
```

Результат работы скрипта dol_hi2.py

Введите объём выборок долей: 1000.
Введите количество словоформ в первой выборке: 200.
Введите количество словоформ во второй выборке: 160.
Введите число степеней свободы: 1.
С вероятностью 0,05 величина критерия будет больше табличного – расчётная величина пси-5.42 больше или равна теоретической табличной 3.84.
Отклонение долей от средней доли –0.18 носит **неслучайный** характер.

Если числовое значение «хи-квадрат» применим для двух степеней свободы, показания обеих формул совпадают. Воспользуемся скриптом **dol_hi2.py**.

Результат работы скрипта dol_hi2.py

Введите объём выборок долей: 1000.
Введите количество словоформ в первой выборке: 200.
Введите количество словоформ во второй выборке: 160.
Введите число степеней свободы: 2.
С вероятностью 0,05 величина критерия будет больше табличного – расчётная величина пси-5.42 меньше теоретической табличной 5.95.
Отклонение долей от средней доли – 0.18 – носит **случайный характер.**

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 достоверные 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 ме-

няющиеся 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 проявлением которого и оказываются наблюдаемые выборочные частоты с их колебаниями.

Контрольные вопросы

1. Дайте определение статистическому термину «доля».

2. Напишите формулу для вычисления доли.

3. Чем отличается формула вероятности события от формулы для определения доли?

4. По какому закону и как изменяются доли словоупотреблений относительно их средних значений?

5. Как определить долю остальных словоупотреблений при известной доле заданных словоупотреблений?

6. При каких условиях для долей позволительно вычислять квадратичное отклонение доли.

7. Напишите формулу квадратичного отклонения доли.

8. В чем формула квадратичного отклонения доли подобна формуле квадратичного отклонения частоты?

9. Для решения каких статистических задач формула квадратичного отклонения долей получает следующий вид:

$$\varepsilon_{1,2} = \sqrt{\overline{p} \cdot \overline{q} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}\,?$$

10. Какие параметры входят в следующую формулу для квадратичного отклонения доли:

$$\varepsilon_{1,2} = \sqrt{\overline{p} \cdot \overline{q} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}?$$

11. Найдите средние доли $\overline{p}$ и $\overline{q}$ при следующих условиях: предположим, были взяты две текстовые выборки, каждая длиной в 1000 знаменательных слов; в первой выборке оказалось 180 глаголов, во второй – 250.

12. Рассчитайте величину квадратичного отклонения средней доли двух сравниваемых совокупностей при следующих условиях: предположим, были взяты две текстовые выборки, каждая длиной в 500 знаменательных слов; в первой выборке оказалось 90 глаголов, во второй – 125.

13. Какую формулу Ван дер Варден предложил для сравнения долей (вероятностей) для критерия «хи-квадрат»?

14. Докажите, что формулы для критерия «хи-квадрат» $\chi^2 = \frac{(x_1 - n_1 \cdot \overline{p})^2}{n_1 \cdot \overline{p} \cdot \overline{q}} + \frac{(x_2 - n_2 \cdot \overline{p})^2}{n_2 \cdot \overline{p} \cdot \overline{q}}$ и $\chi^2 = \frac{(x_1 - \overline{x})^2}{\overline{x} \cdot \overline{q}} + \frac{(x_2 - \overline{x})^2}{\overline{x} \cdot \overline{q}}$ дают одинаковые результаты при равных объёмах выборок $n_1 = n_2$.

15. Почему при вычислении критерия «хи-квадрат» по формуле $\chi^2 = \frac{(x_1 - n_1 \cdot \overline{p})^2}{n_1 \cdot \overline{p} \cdot \overline{q}} + \frac{(x_2 - n_2 \cdot \overline{p})^2}{n_2 \cdot \overline{p} \cdot \overline{q}}$ нужно принимать степень свободы не 1, как советует Ван дер Варден [2], а 2.

Практическое занятие № 4

Задание: Исследовать скрипты **dol.py,dol_hi2.py** по общей методике исследования скриптов Python (Практическое задание № 2, пример). При изменении исходных данных там, где программа определяет характер рассеивания частот возле средней частоты подбором исходных данных добивайтесь противоположного по отношению к шаблону результата. Для каждого такого случая дайте детальное пояснение.

ТЕМА 5. 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 показателей речи. 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 оценками колеблемости, различием и их величины и их статистического характера, их вероятностной специфики, вероятностного качества. Одним словом, возникают 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.

Пример 5.1. Из текстов писателя Л. было взято 10 выборок по 500 словоупотреблений знаменательных слов. Из серии текстов писателя В. было сделано столько же выборок такого же объема. Интуитивно все выборки писателя Л. были определены как более или менее однородные, то же самое можно сказать о выборках из текстов писателя В.

Получены такие числовые данные, характеризующие частоту имен прилагательных: писатель Л.: 72, 65, 78, 71, 70, 74, 80, 90, 68, 82; писатель В.: 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:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{1,2}} \cdot \sqrt{\frac{k_1 \cdot k_2}{k_1 + k_2}},$$

где $\bar{x}_1$ и $\bar{x}_2$ – сравниваемые средние частоты; k_1 и k_2 – число выборок (наблюдений) в двух различных сериях; $s_{1,2}$ – несмещенная оценка среднего квадратичного отклонения в двух сериях выборок, вычисляемая для сравнения двух средних частот по формуле:

$$s_{1,2} = \sqrt{\frac{\sum \left(x_{i1} - \bar{x}_1\right)^2 + \sum \left(x_{i2} - \bar{x}_2\right)^2}{k_1 + k_2 - 2}},$$

где x_{i1} и x_{i2} – наблюдаемые в первой и второй сериях выборок частоты; остальные символы уже хорошо знакомы.

Вернемся к описанной выше задаче. Частоты первого выборочного ряда (из текстов писателя Л.) дают среднюю частоту, равную 75; частоты второго выборочного ряда (из текстов писателя В.) 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 величи-

ну в первую формулу: $t = \frac{82 - 75}{7,5} \cdot \sqrt{\frac{10 \cdot 10}{10 + 10}} = 2,1$. 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 величиной t в 18-й строке таблицы 5.1 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). Для анализа и проверки сделанных выводов создадим скрипт **stud.py.**

Текст скрипта stud.py

```
a=[72, 65, 78, 71, 70, 74, 80, 90, 68, 82]
s = 0
for x in a:
        s =s+x
print('Сумма элементов массива частот выборки 1 – '+str(s))
r=round(s/len(a),0)
k=len(a)
print('Объёмвыборки1 – '+str(k))
print('Средняя частота выборки1 – '+str(r))
s = 0
for x in a:
        s =s+ (x – r) ** 2
d= round(s,0)
print('Сумма квадратов отклонений от средней частот выборки1 –
'+str(d))
b=[80, 93, 84,83, 78, 67, 85, 86, 75, 89]
s1= 0
for x in b:
```

```
        s1 =s1+x # накопления суммы элементов массива частот
print('Сумма элементов массива частот выборки 2 – '+str(s1))
r1=round(s1/len(b),0)
k1=len(b)
print('Объёмвыборки1 – '+str(k1))
print('Средняя частота выборки2 – '+str(r1))
s1= 0
for x in b:
        s1 =s1+ (x – r1) ** 2
d1= round(s1,0)
print('Сумма квадратов отклонений от средней частот выборки1 –
'+str(d1))
s12=round(((d+d1)/(k+k1-2))**0.5,2)
t=round(abs(((r-r1)/s12)*(k*k1/(k+k1))**0.5),2)
e=k+k1-3
c=[12.706,4.303,3.132,2.776,2.571,2.447,2.365,2.306,2.262,2.228,2.20
1,2.179,2.101,2.145,2.131,2.12,2.11,2.101,2.093,2.086,2.08,2.074,2.06
9,2.064,2.06]
c1=float(c[e])
if t<c1:
        print(' Для степени свободы- '+str(e)+ ', расчётная величина
Критерия Стьюдента -' +str(t) +' меньше его\n\
теоретического табличного значения – '+str(c1)+'. Расхождение
средних частот в двух выборках – СЛУЧАЙНОЕ \n')
if t>=c1:
        print(' Для степени свободы- '+str(e)+ ', расчётная величина
Критерия Стьюдента -' +str(t) +' больше его\n\
теоретического табличного значения – '+str(c1)+'. Расхождение ча-
стот в двух выборках – НЕ СЛУЧАЙНОЕ \n')
```

Результат работы скрипта stud.py

```
Сумма элементов массива частот выборки 1 – 750.
Объём выборки 1 – 10.
Средняя частота выборки 1 – 75.0.
Сумма квадратов отклонений от средней частот выборки 1 – 508.0.
Сумма элементов массива частот выборки 2 – 820.
Объём выборки 1 – 10.
Средняя частота выборки 2 – 82.0.
```

Сумма квадратов отклонений от средней частот выборки 1 – 494.0. Для степени свободы – 17, расчётная величина Критерия Стьюдента – 2.1 меньше его теоритического табличного значения – 2.101. Расхождение средних частот в двух выборках – СЛУЧАЙНОЕ.

Таблица 5.1

Числовые значения t

Число степеней свободы	Вероятность большего значения				
	0,5 (50%)	0,2 (20%)	0,1 (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	3,132	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	2,11O	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	1,731	2,064	2,391
25	0,684	1,316	1,708	2,060	2,385

Пример 5.2. 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:

$$\varepsilon_{1,2} = \sqrt{\frac{\sigma_1^2}{k_1} + \frac{\sigma_2^2}{k_2}},$$

где σ_1^2 и σ_2^2 – дисперсии двух серий выборок, средние частоты которых сравниваются (вспомним формулу для вычисления дисперсии: $\sigma = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2}{k}}$; k_1 и k_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. 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. Для анализа и проверки сделанных выводов создадим скрипт **e_p1_p2.py.**

Текст скрипта e_p1_p2.py

```
a=[72, 65, 78, 71, 70, 74, 80, 90, 68, 82]
s = 0
for x in a:
        s =s+x # накопления суммы элементов массива частот
print('Сумма элементов массива частот выборки 1 – '+str(s))
r=round(s/len(a),0) # len(a) – число элементов, round( ,0) округление до целых
```

```
k=len(a)
print('Объёмвыборки1 – '+str(k))
print('Средняя частота выборки1 – '+str(r))
s = 0
for x in a:
        s =s+ (x – r) ** 2 # накопление суммы квадратов отклонения текущих частот от средней
d= round((s/k)**0.5,2)
print('Среднее квадратичное отклонение по выборке 1 – '+str(d))
b=[80, 93, 84,83, 78, 67, 85, 86, 75, 89]
s1= 0
for x in b:
        s1 =s1+x # накопления суммы элементов массива частот
print('Сумма элементов массива частот выборки 2 – '+str(s1))
r1=round(s1/len(b),0) # len(a) -число элементов в массиве частот
k1=len(b)
print('Объём выборки1 – '+str(k1))
print('Средняя частота выборки2 – '+str(r1))
s1= 0
for x in b:
        s1 =s1+ (x – r1) ** 2 # накопление суммы квадратов отклонения текущих частот от средней
d1= round((s1/k1)**0.5,2)
print('Среднее квадратичное отклонение по выборке 2 – '+str(d1))
e12=round(((d**2/k)+(d1**2/k1))**0.5,2)
c1=abs(r-r1)
if 3*e12<c1:
print(' Утроенное квадратичное отклонение разности, -' +str(3*e12) +' меньше разности средних частот – '+str(c1)+'.\n\
Расхождение средних частот в двух выборках- НЕ СЛУЧАЙНОЕ \n')
if 3*e12>=c1:
        print(' Утроенное квадратичное отклонение разности, -' +str(3*e12) +' больше разности средних частот – '+str(c1)+'.\n\
Расхождение средних частот в двух выборках- СЛУЧАЙНОЕ \n').
```

Результаты работы скрипта e_p1_p2.py

Сумма элементов массива частот выборки 1 – 750.

Объём выборки 1 – 10.
Средняя частота выборки 1 – 75.0.
Среднее квадратичное отклонение по выборке 1 – 7.13.
Сумма элементов массива частот выборки 2 – 820.
Объём выборки 1 – 10.
Средняя частота выборки 2 – 82.0.
Среднее квадратичное отклонение по выборке 2 – 7.03.
Утроенное квадратичное отклонение разности, – 9.51 больше разности средних частот – 7.0..
Расхождение средних частот в двух выборках – СЛУЧАЙНОЕ.

Видимо, можно (хотя работы по статистике об этом обычно н не говорят) использовать для сравнения двух выборочных средних частот и «интервалы действительных средних», вычисляемые с помощью формулы ошибки наблюдения (т. е. формулы $L = \frac{t \cdot \sigma}{\sqrt{k}}$ или $L = \frac{t \cdot s}{\sqrt{k}}$). 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, судят по частотам нескольких выборок, о которых и дает некоторое (тоже приближенное) представление формула $L = \frac{t \cdot \sigma}{\sqrt{k}}$. Она показывает те пределы, за которые выход средних частот при повторном изучении текста маловероятен, а значит, действительная средняя всего текста должна лежать в этих именно пределах.

Пример 5.3. 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 $\sigma = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \overline{x})^2}{k}}$, 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 $L = \frac{t \cdot \sigma}{\sqrt{k}}$, 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 $x_{\bar{O}_1}$ и $x_{\bar{O}_2}$. Первый интервал $x_{\bar{O}_1}$ 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 ± 5,1 = 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. Для анализа и проверки сделанных выводов создадим скрипт **d1_d2.py.**

Текст скрипта d1_d2.py

```
a=[72, 65, 78, 71, 70, 74, 80, 90, 68, 82]
c=[12.706,4.303,3.132,2.776,2.571,2.447,2.365,2.306,2.262,2.228,2.20
1,2.179,2.101,2.145,2.131,2.12,2.11,2.101,2.093,2.086,2.08,2.074,2.06
9,2.064,2.06]
s = 0
for x in a:
        s =s+x
print('Сумма элементов массива частот выборки A – '+str(s))
r=round(s/len(a),2)
```

```
k=len(a)
print('Объём выборки A- '+str(k))
print('Средняя частота выборки A – '+str(r))
s = 0
for x in a:
        s =s+ (x – r) ** 2
d= round((s/k)**0.5,1)
print('Сумма квадратов отклонений от средней частот выборки A – '+str(d))
t=c[k-2]
l=round(t*d/(k**0.5),1)
b=[80, 93, 84,83, 78, 67, 85, 86, 75, 89]
s1= 0
for x in b:
        s1 =s1+x
print('Сумма элементов массива частот выборки B – '+str(s1))
r1=round(s1/len(b),0)
k1=len(b)
print('Объём выборки B – '+str(k1))
print('Средняя частота выборки B – '+str(r1))
s1= 0
for x in b:
        s1 =s1+ (x – r1) ** 2
d1= round((s1/k1)**0.5,1)
print('Сумма квадратов отклонений от средней частот выборки B – '+str(d1))
t1=c[k1-2]
l1=round(t1*d1/(k1**0.5),1)
x1=round(r-l,0)
x2=round(r+l,0)
x3=round(r1-l1,1)
x4=round(r1+l1,1)
if (x1<=x3<x2) or (x3<=x1<x4):
        s='Диапазоны погрешностей средних частот пересекаются, отклонение средних частот СЛУЧАЙНО'
elif(x1<=x3<x2) and (x3<=x1<x4):
        s='Диапазоны погрешностей средних частот не пересекаются, отклонение средних частот НЕ СЛУЧАЙНО'
```

print('Оценочный анализ принадлежности массивов частот к одной генеральной совокупности при помощи \n\
определения погрешностей средней частоты каждого массива: '+str(l)+ ' , ' +str(l1)+' с последующим \n\
определением диапазонов изменения средних частот: ' +str(x1)+ '---'+str(x2)+' , '+str(x3)+'---'+str(x4)+' и выводом \n-'+s+'\n')

Результаты работы скрипта d1_d2.py

Сумма элементов массива частот выборки A – 750.
Объём выборки A – 10.
Средняя частота выборки A – 75.0.
Сумма квадратов отклонений от средней частот выборки A – 7.1.
Сумма элементов массива частот выборки B – 820.
Объём выборки B – 10.
Средняя частота выборки B – 82.0.
Сумма квадратов отклонений от средней частот выборки B – 7.0.
Оценочный анализ принадлежности массивов частот к одной генеральной совокупности при помощи
определения погрешностей средней частоты каждого массива: 5.1, 5.0 с последующим
определением диапазонов изменения средних частот: 70.0–80.0 , 77.0–87.0 и выводом:
– диапазоны погрешностей средних частот пересекаются, отклонение средних частот СЛУЧАЙНО.

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 расхо-

ждения частотных рядов случайны и оба ряда рождены одной и той же вероятностью? 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).

Пример 5.4. Как уже было сказано, математическая статистика имеет специальный и очень неплохо действующий инструмент для сравнения двух частотных рядов – «хи-критерий», который 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 $\psi\left(\frac{r_i}{n+1}\right)$. 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 (табл. 5.2). Если в опыте сумма числовых значений «хи», соответствующая порядковым местам частот одной серии в общем ранжированном ряду, превзойдет указанное во вспомогательной таблице 5.3 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. Частоты В): 67, 75, 78, 80, 83, 84, 85, 86, 89, 93. Частоты C=sort(A+B).: C= [65, 67, 68, 70, 71, 72, 74, 75, 78, 78, 80, 80, 82, 83, 84, 85, 86, 89, 90, 93].

Таблица 5.2

Числовые значения $\psi\left(\frac{r_i}{n+1}\right)$

Порядковый номер частоты	n-общее число частот в двух 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

Таблица 5.3

Критические числовые значения суммы $\sum_{i=1}^{n}\psi\left(\frac{r_i}{n+1}\right)$

Разность числа частот в двух выборочных рядах	n-общее число частот в двух 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

Для анализа и проверки нулевой гипотезы создадим скрипт **A_B.py.**

Текст скриптаA_B.py

```
a=[65, 68, 70, 71, 72, 74, 78, 80, 82, 90]
b=[67, 75, 78, 80, 83, 84, 85, 86, 89, 93]
c=a+b
c.sort(reverse=False)
print('Ранжированный ряд суммы A+B – ',c)
l=-1
w=-1
k=-1
m=[]
q=[]
for i in c:
        l+=1
        for j in a:
                if i==j :
                        if c[l]!=c[l-1]:
                                w+=1
                                m=m+[l]
                        if c[l]==c[l-1]:
                                k+=1
                                if k==0:
                                        m[w]=l-1
                                elif k==1:
                                        m[w]=l
                                ifk==2:
k=k-2
d=[-1.66,-1.31,-1.07,-0.88,-0.71,-0.57,-0.43,-0.31,-0.18,-0.06, 0.06,
0.18, 0.31, 0.43, 0.57, 0.71, 0.88, 1.07, 1.31, 1.66]
e=0
print('Ряд позиций ряда A в C -',m)
for i in m:
        e=e+d[i]
r=round(e,2)
print('Расчётное значение суммы весовых коэффициентов позиций
ряда A -',r)
```

```
f=[2.60, 3,11, 3.63, 3.86, 4.08]
z=len(c)-16
x=f[z]
print('Табличное значение предельной суммы весовых коэффициентов позиций ряда A -',x)
if x>=r:
        print('Предположение о несущественности расхождений \n\
между двумя частотными рядами, можно принять.')
elifx<r:
        print('Предположение о несущественности расхождений \n\
между двумя частотными рядами, следует отвергнуть.')
```

Результаты работы скрипта A_B.py при определении порядковых номеров членов ряда A) в ранжированном по возрастанию ряде C) с выводом о характере расхождения между двумя частотными рядами A) и B)

Ранжированный ряд суммы A+B – [65, 67, 68, 70, 71, 72, 74, 75, 78, 78, 80, 80, 82, 83, 84, 85, 86, 89, 90, 93].
Ряд позиций ряда A в C – [0, 2, 3, 4, 5, 6, 8, 11, 12, 18].
Расчётное значение суммы весовых коэффициентов позиций ряда A – -3.7.
Табличное значение предельной суммы весовых коэффициентов позиций ряда A – 3.86.
Предположение о несущественности расхождений
между двумя частотными рядами можно принять.

Результаты работы скрипта A_B.py при определении порядковых номеров членов ряда B) в ранжированном по возрастанию ряде C) с выводом о характере расхождения между двумя частотными рядами A) и B)

Ранжированный ряд суммы A+B – [65, 67, 68, 70, 71, 72, 74, 75, 78, 78, 80, 80, 82, 83, 84, 85, 86, 89, 90, 93].
Ряд позиций ряда B в C – [1, 7, 8, 11, 13, 14, 15, 16, 17, 19].
Расчётное значение суммы весовых коэффициентов позиций ряда B – 3.7.

Табличное значение предельной суммы весовых коэффициентов позиций ряда B – 3.86.
Предположение о несущественности расхождений между двумя частотными рядами можно принять.

По результатам работы скрипта A_B.py можно построить табл. 5.4, с использованием которой поэтапно объясниться работа программы.

Таблица 5.4

Место членов ряда A) и B) в ранжированном ряде C)

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	–	–
Место	12	13	14	15	16	17	18	19	20		
Ряд A	80	82	–	–	–	–	–	90	–		
Ряд B	80	–	83	84	85	86	89	–	93		

В табл. 5.4 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 табл. 5.2 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 3,70, критическое же значение суммы (вспом. табл. 5.3 – 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.

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 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 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 исследования приёмами 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.

Контрольные вопросы

1. В каких случаях лингвист может использовать сравнение средних выборочных частот помимо сравнения наблюдаемых выборочных частот и долей?

2. Чем в лингвистической статистике характеризуются типы и стили речи?

3. Получены числовые данные, характеризующие частоту имен прилагательных: писатель А.: 72, 65, 78, 71, 70, 74, 80, 90, 68, 82; писатель В.: 80, 93, 84,83, 78, 67, 85, 86, 75, 89; что нужно узнать исследователю?

4. Приведите формулу критерия Стьюдента.

5. Для чего вычисляется несмещенная оценка среднего квадратичного отклонения в двух сериях выборок?

6. Приведите формулу для несмещенной оценки среднего квадратичного отклонения в двух сериях выборок.

7. Чему равно число степеней свободы при определении табличного значения коэффициента Стьюдента, если в одной серии 10 значений частот, а в другой – 12?

8. Какой табличный критерий Стьюдента более надёжно оценивает гипотезу о несущественности расхождения частот, при 95% вероятности его больших значений или при 5%?

9. Когда расчётная величина критерия Стьюдента меньше его теоретического табличного значения?

10. Когда расчётная величина Критерия Стьюдента больше его теоритического табличного значения.

11. Какая формула рекомендуется для вычисления так называемого квадратичного отклонения разности средних частот в двух выборках?

12. При каких условиях гипотеза о несущественности расхождения частот отвергается, если для этого сравнивать квадратичное отклонение разности средних частот в двух выборках и саму их разность?

13. При каких условиях гипотеза о несущественности расхождения частот принимается, если для этого сравнивать квадратичное отклонение разности средних частот в двух выборках и саму их разность?

14. Как можно использовать для сравнения двух выборочных средних частот «интервалы действительных средних»?

15. Приведите формулу для «интервалов действительных средних».

16. Что означает для гипотезы о несущественности расхождения частот перекрытие «интервалов действительных средних»?

17. Приведите условие перекрытия интервалов действительных средних.

18. Как судят о тех ошибках, которые мы могли допустить в оценке величины действительной средней?

19. Что показывает табличный коэффициент Стьюдента в формуле для определения ошибки в вычислении действительного среднего?

20. Какой из способов сравнения средних частот в двух выборках проще?

21. Какой из способов сравнения средних частот в двух выборках менее надёжен и почему?

22. Какой способ сравнения двух частотных рядов носит название «хи-критерий» и обозначается большой греческой буквой «хи» – X? Поясните суть способа.

23. Как определяется «хи-критерий» (не путать с «пси-критерием»)?

24. Как влияет на величину расчётного «хи-критерия» смена ранжирования суммарного ряда, например, с убывания на возрастание?

Практическое занятие №5

Задание: Исследовать скрипты **stud.py, e_p1_p2.py, d1_d2.py, A_B.py** по общей методике исследования скриптов **Python** (Практическое задание № 2, пример). При изменении исходных данных там, где программа определяет характер рассеивания частот возле средней частоты подбором исходных данных добивайтесь противоположного по отношению к шаблону результата. Для каждого такого случая дайте детальное пояснение.

ТЕМА 6. 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 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 неизлучавшиеся 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. Выбо-

рочное статистическое исследование как раз и имеет целью познание законов целого на основе изучения его нескольких частей. Это очень важно прежде всего 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 неизученные участки текста однородны изученным. Об ошибке наблюдения средней частоты однажды уже было сказано, и она была показана в действии. Вспомним ее несложную формулу: $L = \frac{t \cdot \sigma}{\sqrt{k}}$, где t – табличный, теоретический коэффициент, величина которого зависит от числа степеней свободы (т. е. для наблюдателя-лингвиста от количества выборок), σ – среднее квадратичное отклонение (или еще лучше пользоваться величиной s – несмещенной оценкой среднего квадратичного отклонения); k – число наблюдений (выборок). Будем думать, что мы уже умеем вычислять среднее квадратичное отклонение (или его несмещенную

оценку); на всякий случай напоминаю одну формулу: $s = \sqrt{\frac{\left(x_i - \overline{\overline{x}}\right)}{k-1}}$.
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 (табл. 6.1).

Таблица 6.1

Числовые значения *t*

Количество выборок	Надёжность 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	3t50	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	2,9S	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

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 и т. д. Чем больше коэффициент, тем надежнее результат, т. е. тем вероятнее определяются 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 величи-

ну, 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.

Пример 6.1. 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 отклонении 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 неисследованный текст однороден, по интересующим нас языковым признакам, исследованным 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 (85 + 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 неисследованных кусках текста (или текстов), однородных исследованному, 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».

Пример 6.2. 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 несмещённые 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% (5 : 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 сравнивать 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: $\delta = \frac{t \cdot \sigma}{\bar{x} \cdot \sqrt{k}}$; 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 $L_p = \frac{2 \cdot \sqrt{p \cdot q}}{\sqrt{n}}$; б) относительная ошибка доли $\delta_p = \frac{2 \cdot \sqrt{q}}{\sqrt{p \cdot n}}$ 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).

Пример 6.3. Допустим, мы имеем выборку длиной в 10 000 (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: $L_p = \frac{2 \cdot \sqrt{0,35 \cdot 0,65}}{\sqrt{10000}} = 0,0093.$ 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: $\delta_p = \frac{2 \cdot \sqrt{q}}{\sqrt{p \cdot n}}$.

Пример 6.4. Относительная ошибка доли должна пониматься аналогично относительной ошибке частоты. Относительная ошибка определении доли – это отношение абсолютной ошибки к величине выборочной доли.

Решение 1: По формуле $\delta_p = \frac{2 \cdot \sqrt{0,65}}{\sqrt{0,35 \cdot 10000}} = 0,027.$

Решение 2: 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.

Сравнение решений. 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 получим приблизительно тот же результат (правда, чуть меньше – 0,027 вместо 0,029; это объясняется тем, что во втором 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 – все-

го 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.

Пример 6.5. 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 $\delta = \frac{2 \cdot \sigma}{\bar{x} \cdot \sqrt{k}}$ можно получить, путем преобразования формулу для определения числа наблюдений (выборок): $k = \frac{4 \cdot \sigma^2}{\bar{x}^2 \cdot \delta^2}$. Введем в эту формулу данные из условия задачи, т. е. $\sigma = 16,5$, $\delta = 0,05$, $\bar{x} = 90$. Получаем:

$$k = \frac{4 \cdot 16,5^2}{90^2 \cdot 0,05^2} = 49.$$

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?

$k = \frac{4 \cdot 7,5^2}{90^2 \cdot 0,05^2} = 11$. 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.

$k = \frac{4 \cdot 16,5^2}{90^2 \cdot 0,1^2} = 12.$ И в этом случае число выборок заметно падает.

И еще один вариант: допустим, что средняя часта глаголов была не 90, а 110, остальные условия задачи остались как в первоначальном ее варианте. Что покажет формула?

$k = \frac{4 \cdot 16,5^2}{190^2 \cdot 0,05^2} = 32$. 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: $k = \frac{4 \cdot \sigma^2}{L^2}$ 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, $k = \frac{4 \cdot 5,5^2}{5.0^2} = 6$. Оказывается, нужно всего 6 выборок, чтобы по-

лучить среднюю с ошибкой, не превышающей 5 единиц. Правда, в нашу формулу молчаливо введен жесткий коэффициент, равный 2 ($L = \frac{2 \cdot \sigma}{\sqrt{k}}$, именно эта формула преобразована в формулу, $k = \frac{4 \cdot \sigma^2}{L^2}$) но мы помним, что надежность определения средней в заданных ошибкой пределах зависит от коэффициента *t* и что он изменчив – 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: $n \approx \frac{4 \cdot p \cdot q}{L^2}$;

б) формула для определения объема выборки по заданной относительной ошибке доли: $n = \frac{4 \cdot q}{\delta^2 \cdot p}$.

Применим первую и вторую формулы к решению конкретных задач.

Пример 6.6. 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?

$$n = \frac{4 \cdot pq}{L^2} = \frac{4 \cdot 0,07 \cdot 0,93}{0,005^2} = 10416 .$$

Пример 6.7. Доля наречий та же. Экспериментатор хочет определить ее с относительной ошибкой, не превышающей 0,05. Какой должна быть длина выборки?

$n = \frac{4 \cdot q}{\delta^2 \cdot p} = \frac{4 \cdot 0,93}{0,005^2 \cdot 0,07} = 21257$ 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 – 21000.

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 изучавшийся 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.

Контрольные вопросы

1. Почему применение статистических инструментов для изучения языка и речи привлекает внимание лингвистов?
2. Как разграничить понятия «статистическое исследование языка» и «статистическое описание»?
3. Какими недостатками обладает статистическое описание?
4. Какова цель выборочного статистического исследования?
5. В чём состоит отличие средней выборочной частоты языковых явлений от «действительной средней»?
6. Дайте определение термину «ошибку наблюдения» при выборочном статистическом исследовании.
7. Как называются пределы, в которых может находиться «действительная средняя частота» или «действительная доля», если предположить, что неизученные участки текста однородны изученным?
8. Приведите формулу для ошибки наблюдения, поясните влияние входящих в формулу переменных на величину ошибки.
9. В чём особенности применения в формуле для ошибки наблюдения несмещенной оценкой среднего квадратичного отклонения вместо среднего квадратичного отклонения?

10. Как выбирать подходящую величину для коэффициента *t* в формуле ошибки наблюдения?

11. Как изменяется надёжность результата наблюдения с увеличением табличного коэффициента *t* в формуле ошибки наблюдения?

12. Какое значение коэффициента *t* в формуле ошибки наблюдения, лингвисты выбирают тогда, когда не требуется особо большая точность и надежность наблюдений?

13. Что значит 95%-ная надежность коэффициента *t* в формуле ошибки наблюдения?

14. Дайте определение абсолютной и относительной ошибке наблюдения.

15. Приведите формулу для определения относительной ошибки наблюдения частот.

16. Приведите формулы для определения абсолютной и относительной ошибки при наблюдении выборок долей, укажите, как влияют переменные входящие в эти формулы на ошибки наблюдения.

17. Что позволяют планировать при статистических исследованиях языковых явлений формулы абсолютной и относительной ошибки средней частоты и доли?

18. Из формулы относительной ошибки частоты $\delta = \frac{2 \cdot \sigma}{\overline{x} \cdot \sqrt{k}}$ путем преобразования получите формулу для определения числа наблюдений (выборок).

19. Приведите формулу для определения числа выборок по известной из предшествующего опыта величине среднего квадратичного отклонения и планируемой абсолютной ошибке.

20. В чём разница между числом выборок и объёмом выборки?

21. Приведите формулу для определения объема выборки по заданной абсолютной ошибке доли.

22. Приведите формулу для определения объема выборки по заданной относительной ошибке доли.

Практическое занятие № 6

Задание: На Python написать скрипты для решения примеров 6.1−6.7. Проверить работу скриптов по числовым данным, представленным в примерах. В Word составить отчёт, содержащий текст и результаты работы скриптов.

ТЕМА 7. ОРГАНИЗАЦИЯ СТАТИСТИЧЕСКОГО ИЗУЧЕНИЯ ЯЗЫКА И РЕЧИ НА ОСНОВЕ СОВРЕМЕННЫХ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Почему Python? Python – простой, но все же сильный язык программирования с превосходной функциональностью для обработки лингвистических данных. Python может быть загружен бесплатно. Инсталляторы доступны для всех платформ.

Решение задач обработки текстов на естественном языке предполагает использование больших объемов лингвистических данных, или, другими словами, предполагает работу с корпусами текстов. Изучение данной темы поможет лингвисту найти ответ на следующие вопросы: какие известны корпуса текстов и лексические ресурсы и как получить к ним доступ, используя Python; какие полезные конструкции имеет Python для статистического изучения языка и речи.

Изучение темы основано на языке программирования Python вместе с общедоступной библиотекой, названной **Набором инструментов естественного языка (NLTK).** NLTK включает обширное программное обеспечение, данные и документацию, все свободно загружается с http://www .nltk.org/.

Изучение темы не требует предварительных знаний по программированию на Python. Все приведенные в данной, как и в предыдущих темах, примеры легко загружаются и выполняются через графический интерфейс Python.

Прежде чем приступить к работе с темой следует установить необходимые коллекции корпусов текстов. Для этого в командную строку Python вводим следующие команды:

```
>>> import nltk
>>> nltk.download()
```

Появляется окно загрузчика корпусов текста.

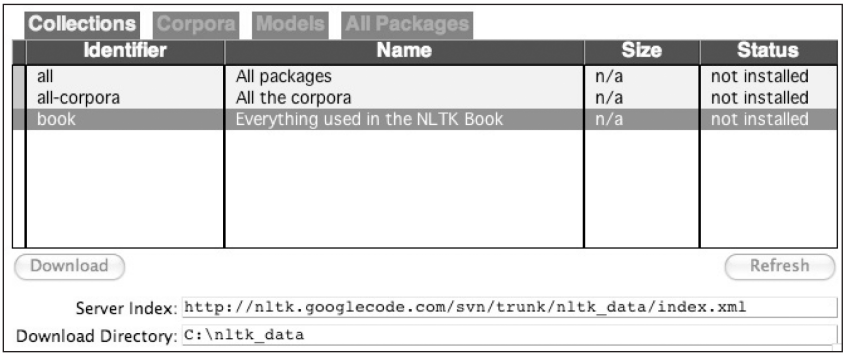

Рис. 7.1. Загрузка книжной коллекции NLTK

Вкладка **Collections** на загрузчике показывает, как пакеты сгруппированы в наборы, и Вы должны выбрать **book**, чтобы получить все данные, требуемые для примеров и упражнений по этой теме. Коллекция состоит приблизительно из 30 сжатых файлов, требующих дискового пространства приблизительно 100 МБ. Полная коллекция данных (т. е., **все** в загрузчике) превышает приблизительно в пять раз указанный объём и продолжает расширяться.

Теперь можно приступить к изучению возможностей NLTK по статистическому анализу текста. Для автоматического определения слов, которые являются наиболее информативными для текстов определенного жанра или определенной тематики сначала необходимо построить частотный список или частотное распределение.

Частотное распределение указывает на частоту, с которой в тексте встречается каждое из слов. Такой частотный список называют распределением потому, что он указывает каким образом общее количество слов распределяется между словарными статьями (оригинальные слова) в тексте.

Учитывая, что построение частотных распределений часто необходимо при обработке естественного языка в NLTK реализован отдельный класс **FreqDist** в модуле **nltk.probability**.

1. Класс FreqDist для простых статистических исследований

```
*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
```

Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908

Применим этот класс для нахождения 50 наиболее частотных слов в тексте Moby Dick.

>>>**fdist=FreqDist(text1)** #Название текста указывается как аргумент класса
>>>**fdist** # Подсчет общего количества слов
<FreqDistwith 260819 outcomes>
>>>**vocab=fdist.keys()**#Установка списка оригинальных слов текста и сортировка их по количеству в тексте
>>>list(vocab)[:50] #Просмотр первых 50 наиболее употребляемых слов
[',', 'the', '.', 'of', 'and', 'a', 'to', ';', 'in', 'that', '"', '-', 'his', 'it', 'I', 's', 'is', 'he', 'with', 'was', 'as', '"', 'all', 'for', 'this', '!', 'at', 'by', 'but', 'not', '--', 'him', 'from', 'be', 'on', 'so', 'whale', 'one', 'you', 'had', 'have', 'there', 'But', 'or', 'were', 'now', 'which', '?', 'me', 'like']
>>> fdist['the']
13721
>>> fdist['of']
6536
>>> fdist['with']
1659
>>>fdist['whale']
906
>>>fdist.plot(50)

Среди этих 50 слов только одно дает определенную информацию о тексте (whale) и это слово встречается в тексте 906 раз. Все другие слова является неинформативными или служебными.

Чтобы определить, какую часть текста занимают служебные слова, найдем их суммарную частоту, построив график (рис. 7.2) с помощью fdist.plot (50).

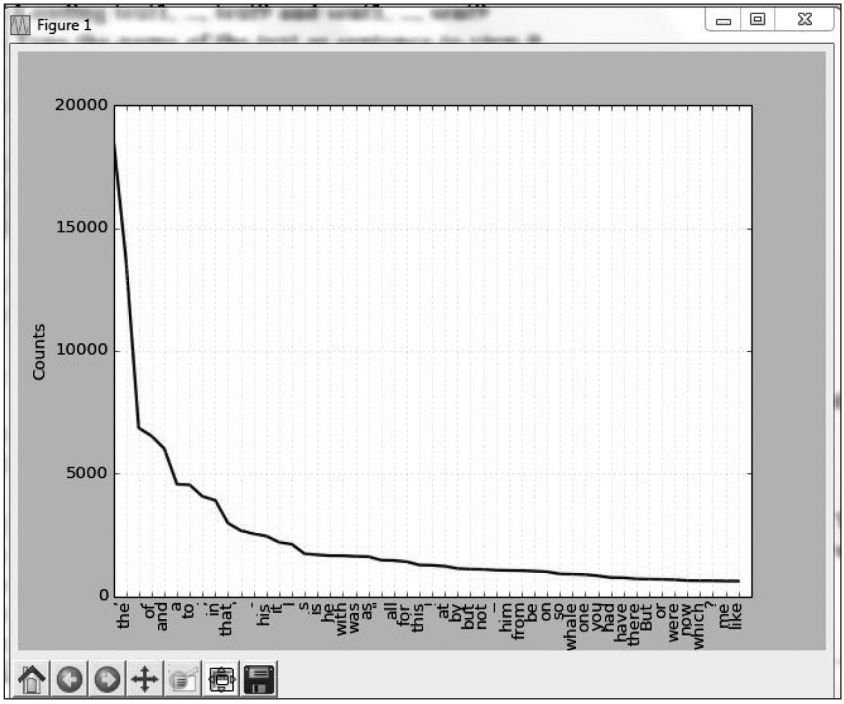

Рис. 7.2. Частотное распределение 50 наиболее употребляемых слов текста: Moby Dickby Herman Melville 1851

Если наиболее частотные слова не дают представления о тексте, то попробуем с помощью fdist.hapaxes () построить список слов, которые встречаются только один раз.

```
>>>fdist=FreqDist(text1)
>>> w=fdist.hapaxes()
>>> len(w)
9002
>>> w[250:300]
['Alarmed', 'Albemarle', 'Albert', 'Albicore', 'Aldrovandi',
'Aldrovandus', 'Alexanders', 'Alfred', 'Algerine', 'Algiers', 'Alike',
'Alive', 'Alleghanian', 'Alleghanies', 'Alley', 'Almanack', 'Almighty',
```

'Ambergriese', 'Ambergris', 'Americas', 'Amittai', 'Anacharsis', 'Anak', 'Anatomist', 'Andrew', 'Angel', 'Angelo', 'Angels', 'Animated', 'Annawon', 'Anno', 'Anomalous', 'Antiochus', 'Antony', 'Antwerp', 'Anvil', 'Anyhow', 'Anyway', 'Apollo', 'Apoplexy', 'Applied', 'Apply', 'Archbishop', 'Arched', 'Archipelagoes', 'Arethusa', 'Argo', 'Arion', 'Arkansas', 'Arkite']

В данном тексте 9002 слова, среди них есть непереводимые – Albemarle, Aldrovandus. Смысл остальных слов без контекста не ясен.

1.1. Выбор слов из текста

Попробуем найти более длинные слова, надеясь на то что они окажутся наиболее информативными.

```
>>> V = set(text1)
>>> long_words = [w for w in V if len(w) > 15]
>>> sorted(long_words)
```

['CIRCUMNAVIGATION', 'Physiognomically', 'apprehensiveness', 'cannibalistically', 'characteristically', 'circumnavigating', 'circumnavigation', 'circumnavigations', 'comprehensiveness', 'hermaphroditical', 'indiscriminately', 'indispensableness', 'irresistibleness','physiognomically','preternaturalness','responsibilities', 'simultaneousness', 'subterraneousness', 'supernaturalness', 'superstitiousness', 'uncomfortableness', 'uncompromisedness', 'undiscriminating', 'uninterpenetratingly']

В приведенном списке слова были отобраны из условия, что их длина больше 15 символов. Однако без учёта частот их упоминания в тексте говорить об их информативности не представляется возможным. Поэтому рассмотрим вариант отбора, лишённый указанного недостатка.

```
>>>f5 = FreqDist(text5)
>>> sorted([w for w in set(text5) if len(w) > 7 and f5[w] > 7])
```

['#14-19teens', '#talkcity_adults', '((((((((((', '.....', 'Question', 'actually', 'anything', 'computer', 'cute.-ass', 'everyone', 'football', 'innocent', 'listening', 'remember', 'seriously', 'something', 'together', 'tomorrow', 'watching']

В приведенном списке слова были отобраны из условия, что их длина больше 7 символов, а каждое слово встречается в тексте не менее 7 раз. Последнее легко проверить, подставив любое слово из списка в f5.

```
>>>f5['watching']
10
```

1.2. Коллокации и биграммы

Сопоставление (collocation) обозначает устойчивую фразеологическую единицу, то есть это словосочетание, которое встречается гораздо чаще, чем его составляющие по отдельности. Например, red wine – это сопоставление, a the wine – нет. Характерной чертой коллокации является то, что они устойчивы к замене одного из слов на другое, подобное по содержанию (maroon wine).

Для того чтобы построить коллокации, сначала нужно построить на основе текста пары слов, или биграммы. Для этого можно использовать функцию bigrams ():

```
>>>list(bigrams(['more', 'is', 'said', 'than', 'done']))
[('more', 'is'), ('is', 'said'), ('said', 'than'), ('than', 'done')]
```

Поскольку коллокации – это частотные биграммы с учетом случаев редких слов, то нам нужно найти такие биграммы, частота которых выше, чем частоты слов, из которых он состоит. Функция collocations () реализует следующие действия.

```
>>>text4.collocations()
Building collocations list
United States; fellow citizens; four years; years ago; Federal
Government; General Government; American people; Vice President; Old
World; Almighty God; Fellow citizens; Chief Magistrate; Chief Justice;
God bless; every citizen; Indian tribes; public debt; one another;
foreign nations; political parties

>>> text8.collocations()
Building collocations list
would like; medium build; social drinker; quiet nights; non smoker;
long term; age open; Would like; easy going; financially secure; fun
times; similar interests; Age open; weekends away; poss rship; well
presented; never married; single mum; permanent relationship; slim
build
```

Коллокации характерны для текстов различных тематик и жанров. Кроме подсчета отдельных слов, интересно также осуществить подсчет длин слов в тексте, используя FreqDist.

```
>>> fdist = FreqDist([len(w) for w in text1])
>>> fdist
<FreqDist with 19 samples and 260819 outcomes>
>>> list(fdist.keys())
```

[3, 1, 4, 2, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20]

Видим, что в тексте встречаются слова различной длины от 1 до 20 символов. Частоту различных длин слов можно посмотреть:

```
>>>fdist.items()
```

[(3, 50223), (1, 47933), (4, 42345), (2, 38513), (5, 26597), (6, 17111), (7, 14399), (8, 9966), (9, 6428), (10, 3528), (11, 1873), (12, 1053), (13, 567), (14, 177), (15, 70), (16, 22), (17, 12), (18, 1), (**20, 1)]**

```
>>>fdist.max()
3
>>>fdist[3]
50223
>>>fdist.freq(3)
0.19255882431878046
```

Чаще встречаются слова с длиной 3 символа, и такие слова составляют около 20% всего текста. Можно осуществить анализ длин слов для текстов различных жанров, авторов и языков.

2. Доступ к корпусам текстов

Корпус текстов это большой набор текстов. Многие корпуса разработаны с сохранением баланса между текстами разных жанров или авторов. При работе с корпусами важно иметь средства доступа как к отдельным текстам, так и к отдельным частям этих текстов а также и к отдельным словам. Для того чтобы получить доступ к корпусам, перед началом работы следует их импортировать.

```
>>> import nltk
>>> from nltk.corpus import *
```

2.1. Корпус Гутенберга

В NLTK входит небольшая часть текстов из электронного архива текстов Project Gutenberg, содержащего 25000 бесплатных электронных книг различных авторов (http://www.gutenberg.org/). Тексты произведений в отдельных файлах. Для получения имен файлов (идентификаторов файлов), в которых хранятся тексты, нужно использовать следующую функцию: fileids () – названия файлов в корпусе.

```
>>> gutenberg.fileids()
['austen-emma.txt', 'austen-persuasion.txt', 'austen-sense.txt', 'bible-
kjv.txt', 'blake-poems.txt', 'bryant-stories.txt', 'burgess-busterbrown.
txt', 'carroll-alice.txt', 'chesterton-ball.txt', 'chesterton-brown.txt',
'chesterton-thursday.txt', 'edgeworth-parents.txt', 'melville-moby_dick.
txt', 'milton-paradise.txt', 'shakespeare-caesar.txt', 'shakespeare-hamlet.
txt', 'shakespeare-macbeth.txt', 'whitman-leaves.txt']
```

Для работы с первым текстом этого корпуса (роман «Эмма», автор Джейн Остин) создаем переменную **emma** и можем найти, сколько слов содержит этот текст.

```
>>> emma = nltk.corpus.gutenberg.words('austen-emma.txt')
>>> len(emma)
192427
```

При создании переменной emma было использовано функцию words () объекта gutenberg пакета corpus библиотеки NLTK. Аналогичный результат можно достичь, используя более компактную запись конструкций Python.

```
>>> from nltk.corpus import gutenberg
>>> gutenberg.fileids()
['austen-emma.txt', 'austen-persuasion.txt', 'austen-sense.txt', 'bible-
kjv.txt', 'blake-poems.txt', 'bryant-stories.txt', 'burgess-busterbrown.
txt', 'carroll-alice.txt', 'chesterton-ball.txt', 'chesterton-brown.txt',
'chesterton-thursday.txt', 'edgeworth-parents.txt', 'melville-moby_dick.
txt', 'milton-paradise.txt', 'shakespeare-caesar.txt', 'shakespeare-hamlet.
txt', 'shakespeare-macbeth.txt', 'whitman-leaves.txt']
```

2.1.1. Конкорданс

Конкорданс – традиционный, давно известный, но до сих пор недостаточно оцененный способ изучения текста. Он предоставляет полный индекс слов в ближайших и расширенных контекстах. Компьютерный конкорданс позволит Вам сравнить все контексты употребления слова, проанализировать их, увидеть слово в самом тексте художественного произведения. Для построения конкорданса слова «**surprize**» в тексте «austen-emma.txt 'необходимо использовать следующие выражения:

```
>>> emma = nltk.Text(nltk.corpus.gutenberg.words('austen-emma.txt'))
>>> emma.concordance(«surprize»)
Building index...
Displaying 25 of 37 matches:
er father , was sometimes taken by surprize at his being still able to pity `
hem do the other any good ." " You surprize me ! Emma must do Harriet good : a
Knightley actually looked red with surprize and displeasure , as he stood up ,
r . Elton , and found to his great surprize , that Mr . Elton was actually on
d aid ." Emma saw Mrs . Weston ' s surprize , and felt that it must be great ,
father was quite taken up with the surprize of so sudden a journey , and his f
y , in all the favouring warmth of surprize and conjecture . She was , moreove
he appeared , to have her share of surprize , introduction , and pleasure . Th
ir plans ; and it was an agreeable surprize to her , therefore , to perceive t
talking aunt had taken me quite by surprize , it must have been the death of m
f all the dialogue which ensued of surprize , and inquiry , and congratulation
 the present . They might chuse to surprize her ." Mrs . Cole had many to agre
the mode of it , the mystery , the surprize , is more like a young woman ' s s
 to her song took her agreeably by surprize -- a second , slightly but correct
" " Oh ! no -- there is nothing to surprize one at all .-- A pretty fortune ;
t to be considered . Emma ' s only surprize was that Jane Fairfax should accep
of your admiration may take you by surprize some day or other ." Mr . Knightle
ation for her will ever take me by surprize .-- I never had a thought of her i
 expected by the best judges , for surprize -- but there was great joy . Mr .
 sound of at first , without great surprize . " So unreasonably early !" she w
d Frank Churchill , with a look of surprize and displeasure .-- " That is easy
; and Emma could imagine with what surprize and mortification she must be retu
tled that Jane should go . Quite a surprize to me ! I had not the least idea !
 . It is impossible to express our surprize . He came to speak to his father o
g engaged !" Emma even jumped with surprize ;-- and , horror – struck , exclai
```

2.1.2. Статистические данные о тексте

Попробуем написать небольшую программу для получения информации о каждом тексте из корпуса. Построим цикл по переменной fileid, которая соответствует идентификатору файла с текстом, и на каждом шагу будем определять некоторую статистическую информацию, которую для компактности записи будем отражать целыми числами int ().

```
>>> import nltk
>>> from nltk.corpus import gutenberg
>>> for fileid in gutenberg.fileids():
    num_chars = len(gutenberg.raw(fileid))
    num_words = len(gutenberg.words(fileid))
    num_sents = len(gutenberg.sents(fileid))
    num_vocab = len(set([w.lower() for w in gutenberg.words(fileid)]))
    print (int(num_chars/num_words), int(num_words/num_sents),
    int(num_words/num_vocab), fileid)
```

```
4 24 26 austen-emma.txt
4 26 16 austen-persuasion.txt
4 28 22 austen-sense.txt
……..
```

Операция деления дает целочисленные результаты (с округлением). Для получения результатов без округления следует использовать тип данных результатов деления **не int(), а float()**, причем количество знаков n после запятой можно установить при помощи **round(,n).**

```
>>> for fileid in gutenberg.fileids():
num_chars = len(gutenberg.raw(fileid))
    num_words = len(gutenberg.words(fileid))
    num_sents = len(gutenberg.sents(fileid))
    num_vocab = len(set([w.lower() for w in gutenberg.words(fileid)]))
    print(round(float(num_chars/num_words),2),
    round(float(num_words/num_sents),2),
    round(float(num_words/num_vocab),2), fileid)
```

```
4.61 24.82 26.2 austen-emma.txt
4.75 26.20 16.82 austen-persuasion.txt
4.75 28.32 22.11 austen-sense.txt
……..
```

Данная программа отображает следующие статистические данные для каждого из текстов: **средняя длина слова; средняя длина предложения; значение лексического разнообразия (отношение общего количества слов с количеством оригинальных слов).** Числовые значения (одинаковые для всех текстов) указывают, что для английского языка среднее значение длины слова составляет 4 символа (**на самом деле 3 поскольку переменная num_chars содержит и пробелы**). В отличие от длины слова следующие числовые значения отличаются и в **некоторой степени характерны для разных авторов.**

В предыдущем примере использовалась функция. raw () для доступа к тексту книги без его разделения на отдельные слова. Эта функция позволяет получить доступ к содержимому файла без какой-либо его предварительной лингвистической обработки. Поэтому использование len (gutenberg.raw ('blake-poems.txt') позволяет установить количество символов (включая пробелы) в тексте. Функция sents () делит текст на отдельные предложения, и каждое предложение представляется как список лент, где ленты – отдельные слова.

```
>>> import nltk
>>> from nltk.corpus import *
>>> macbeth_sentences = gutenberg.sents('shakespeare-macbeth.txt')
>>> macbeth_sentences
[['[', 'The', 'Tragedie', 'of', 'Macbeth', 'by', 'William', 'Shakespeare',
'1603', ']'], ['Actus', 'Primus', '.'], ...]
>>> macbeth_sentences[1037]
['Good', 'night', ',', 'and', 'better', 'health', 'Attend', 'his', 'Maiesty']
>>> longest_len = max([len(s) for s in macbeth_sentences])
>>> [s for s in macbeth_sentences if len(s) == longest_len]

[['Doubtfull', 'it', 'stood', ',', 'As', 'two', 'spent', 'Swimmers', ',', 'that', 'doe',
'cling', 'together', ',', 'And', 'choake', 'their', 'Art', ':', 'The', 'mercilesse',
'Macdonwald', '(', 'Worthie', 'to', 'be', 'a', 'Rebell', ',', 'for', 'to', 'that', 'The',
'multiplying', 'Villanies', 'of', 'Nature', 'Doe', 'swarme', 'vpon', 'him', ')',
'from', 'the', 'Westerne', 'Isles', 'Of', 'Kernes', 'and', 'Gallowgrosses', 'is',
'supply', '"', 'd', ',', 'And', 'Fortune', 'on', 'his', 'damned', 'Quarry', 'smiling',
',', 'Shew', '"', 'd', 'like', 'a', 'Rebells', 'Whore', ':', 'but', 'all', '"', 's', 'too',
'weake', ':', 'For', 'braue', 'Macbeth', '(', 'well', 'hee', 'deserues', 'that',
```

'Name', ')', 'Disdayning', 'Fortune', ',', 'with', 'his', 'brandisht', 'Steele', ',', 'Which', 'smoak', '"', 'd', 'with', 'bloody', 'execution', '(', 'Like', 'Valours', 'Minion', ')', 'caru', '"', 'd', 'out', 'his', 'passage', ',', 'Till', 'hee', 'fac', '"', 'd', 'the', 'Slaue', ':', 'Which', 'neu', '"', 'r', 'shooke', 'hands', ',', 'nor', 'bad', 'farwell', 'to', 'him', ',', 'Till', 'he', 'vnseam', '"', 'd', 'him', 'from', 'the', 'Naue', 'toth', '"', 'Chops', ',', 'And', 'fix', '"', 'd', 'his', 'Head', 'vpon', 'our', 'Battlements']]

2.2. Текст из Интернета [3]

Project Gutenberg включает тысячи книг, и представляет литературный язык. Для работы с менее формальным языком NLTK содержит набор текстов в Интернете: тексты с форума, тексты из фильма «Пираты карибского моря», тексты личных объявлений, телефонные разговоры, обзор вин:

```
>>> from nltk.corpus import webtext
>>> for fileid in webtext.fileids():
        print (fileid, webtext.raw(fileid)[:65])
```

firefox.txt Cookie Manager: «Don't allow sites that set removed cookies to se
grail.txt SCENE 1: [wind] [clop clop clop]
KING ARTHUR: Whoa there! [clop
overheard.txt White guy: So, do you have any plans for this evening? Asian girl
pirates.txt PIRATES OF THE CARRIBEAN: DEAD MAN'S CHEST, by Ted Elliott & Terr
singles.txt 25 SEXY MALE, seeks attrac older single lady, for discreet encoun
wine.txt Lovely delicate, fragrant Rhone wine. Polished leather and strawb

Также в NLTK входит корпус сообщений из чатов, созданный в Naval Postgraduate School для исследований с целью автоматического обнаружения Интернет-преступников. Этот корпус содержит 10000 анонимных сообщений в которых имена пользователей заменены по шаблону «UserNNN», а также удалена другая персональная информация. Корпус организован, 15 отдельных файлов,

каждый из которых содержит несколько сотен сообщений с определенной датой создания и возрастных данных авторов (подростки, 20-ти, 30-ти и 40-ка летние взрослые). Название файла содержит информацию о дате, возрастную группу и количество сообщений, например, файл 10-19-20s_706posts.xml содержит 706 сообщений двадцатилетних авторов от 19 октября 2006 года.

```
>>> from nltk.corpus import nps_chat
>>> chatroom = nps_chat.posts('10-19-20s_706posts.xml')
>>> chatroom[123]
['i', 'do', "n't", 'want', 'hot', 'pics', 'of', 'a', 'female', ',', 'I', 'can', 'look', 'in', 'a', 'mirror', '.']
```

2.3. Корпус Brown

Корпус Brown – это первый корпус английского языка объемом один миллион слов, созданый в 1961–1964 гг. в университете Brown. Этот корпус содержит тексты из 500 различных источников, соответствующих различным жанрам. В табл. 7.1 приведены примеры для каждого из жанров.

Таблица 7.1

Примеры текстов для каждого из жанров корпуса Brown

ID	**Файл**	**Жанр**	**Описание текста**
A16	ca16	news	Chicago Tribune: *Society Reportage*
B02	cb02	editorial	Christian Science Monitor: *Editorials*
C17	cc17	reviews	Time Magazine: *Reviews*
D12	cd12	religion	Underwood: *Probing the Ethics of Realtors*
E36	ce36	hobbies	Norling: *Renting a Car in Europe*
F25	cf25	lore	Boroff: *Jewish Teenage Culture*
G22	cg22	belles_lettres	Reiner: *Coping with Runaway Technology*
H15	ch15	government	US Office of Civil and Defence Mobilization: *The Family Fallout Shelter*
J17	cj19	learned	Mosteller: *Probability with Statistical Applications*
K04	ck04	fiction	W.E.B. Du Bois: *Worlds of Color*
L13	cl13	mystery	Hitchens: *Footsteps in the Night*
M01	cm01	science_fiction	Heinlein: *Stranger in a Strange Land*
N14	cn15	adventure	Field: *Rattlesnake Ridge*
P12	cp12	romance	Callaghan: *A Passion in Rome*
R06	cr06	humor	Thurber: *The Future, If Any, of Comedy*

Используя средства NLTK, можно получить доступ к этому корпусу как в список слов или список предложений (каждое предложение – список слов). Также доступна возможность выбора текстов отдельной категории или отдельного файла.

```
>>> from nltk.corpus import brown
>>> brown.categories()
['adventure', 'belles_lettres', 'editorial', 'fiction', 'government', 'hobbies',
'humor', 'learned', 'lore', 'mystery', 'news', 'religion', 'reviews', 'romance',
'science_fiction']
>>> brown.words(categories='news')
['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', ...]
>>> brown.words(fileids=['cg22'])
['Does', 'our', 'society', 'have', 'a', 'runaway', ',', ...]
>>> brown.sents(categories=['news', 'editorial', 'reviews'])
[['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', 'Friday', 'an',
'investigation', 'of', «Atlanta's», 'recent', 'primary', 'election', 'produced',
'``', 'no', 'evidence', «''», 'that', 'any', 'irregularities', 'took', 'place', '.'],
['The', 'jury', 'further', 'said', 'in', 'term-end', 'presentments', 'that', 'the',
'City', 'Executive', 'Committee', ',', 'which', 'had', 'over-all', 'charge', 'of',
'the', 'election', ',', '``', 'deserves', 'the', 'praise', 'and', 'thanks', 'of', 'the',
'City', 'of', 'Atlanta', «''», 'for', 'the', 'manner', 'in', 'which', 'the', 'election',
'was', 'conducted', '.'], ...]
```

Корпус Brown – удобный ресурс для систематического изучения различий между жанрами или другими словами для исследования стилистики текстов. Попробуем сравнить жанры и установить, каким образом в текстах разных жанров используются модальные глаголы. Для этого нужно сделать подсчеты употребления различных модальных глаголов для разных жанров.

```
>>> import nltk
>>> from nltk.corpus import brown
>>> news_text = brown.words(categories='news')
>>> fdist = nltk.FreqDist([w.lower() for w in news_text])
>>> modals = ['can', 'could', 'may', 'might', 'must', 'will']
>>> for m in modals:
        print(m + ':', fdist[m])
can: 94
```

could: 87
may: 93
might: 38
must: 53
will: 389

2.4. Корпус информационного агентства Рейтер

Корпус Reuters содержит 10788 текстов новостей общим объемом 1,3 миллиона слов. Все тексты разделены на категории по 90 темам и разделены на два набора (тренировки и тестирования). Такое разделение необходимо для тренировки и тестирования алгоритмов автоматического определения тематики текста.

```
>>> import nltk
>>> from nltk.corpus import reuters
>>> reuters.fileids()[:50]
['test/14826', 'test/14828', 'test/14829', 'test/14832', 'test/14833',
'test/14839', 'test/14840', 'test/14841', 'test/14842', 'test/14843',
'test/14844', 'test/14849', 'test/14852', 'test/14854', 'test/14858',
'test/14859', 'test/14860', 'test/14861', …
>>> reuters.categories()
['acq', 'alum', 'barley', 'bop', 'carcass', 'castor-oil', 'cocoa', 'coconut',
'coconut-oil', 'coffee', 'copper', 'copra-cake', 'corn', 'cotton', 'cotton-oil',
'cpi', 'cpu', 'crude', 'dfl', 'dlr', 'dmk', 'earn',……
```

В отличие от корпуса Brown, категории текстов в этом корпусе могут накладываться друг на друга, поскольку тематика новостей (газетных публикаций) преимущественно касается многих тем. Средствами NLTK можно обратиться к темам, которых касаются в одном или нескольких текстах, или наоборот узнать весь перечень текстов, относящихся к определенной категории.

```
>>> reuters.categories('training/9865')
['barley', 'corn', 'grain', 'wheat']
>>> reuters.categories(['training/9865', 'training/9880'])
['barley', 'corn', 'grain', 'money-fx', 'wheat']
>>> reuters.fileids('barley')
```

['test/15618', 'test/15649', 'test/15676', 'test/15728', 'test/15871', 'test/15875', 'test/15952', 'test/17767', 'test/17769', 'test/18024', 'test/18263','test/18908','test/19275','test/19668','training/10175',……..
>>> reuters.fileids(['barley', 'corn'])
['test/14832', 'test/14858', 'test/15033', 'test/15043', 'test/15106', 'test/15287', 'test/15341', 'test/15618', 'test/15648', 'test/15649', 'test/15676', 'test/15686', 'test/15720', 'test/15728', 'test/15845', 'test/15856', 'test/15860', 'test/15863',….

2.5. Корпус инаугурационных приёмов президентов США

Знакомясь с библиотекой программ NLTK, мы работали с одним корпусом и рассматривали весь корпус как один текст, что давало возможность найти место отдельного слова в текстах речей, начиная от первого слова первой речи. На самом деле корпус – это набор 55 текстов, каждый из которых является речью одного президента. Интересная особенность этого корпуса – возможность исследовать распределение текстов по промежутками. Название каждого текста содержит год произнесения речи, и, соответственно есть возможность получить доступ к этой информации. Получим доступ к первым четырем символам названия всех файлов [fileid [4], которые отображают год инаугурации.

>>> import nltk
>>> from nltk.corpus import inaugural
>>> inaugural.fileids()[:10]
['1789-Washington.txt', '1793-Washington.txt', '1797-Adams.txt', '1801-Jefferson.txt', '1805-Jefferson.txt', '1809-Madison.txt', '1813-Madison.txt','1817-Monroe.txt','1821-Monroe.txt','1825-Adams.txt']
>>> [fileid[:4] for fileid in inaugural.fileids()]
['1789', '1793', '1797', '1801', '1805', '1809', '1813', '1817', '1821', '1825', '1829', '1833', '1837', '1841', '1845', '1849', '1853', '1857', '1861', '1865', '1869', '1873', '1877', '1881', '1885', '1889', '1893', '1897', '1901', '1905', '1909', '1913', '1917', '1921', '1925', '1929', '1933', '1937', '1941', '1945', '1949', '1953', '1957', '1961', '1965', '1969', '1973', '1977', '1981', '1985', '1989', '1993', '1997', '2001', '2005', '2009']

2.6. Аннотированные (размечены) корпуса текстов

Большинство корпусов текстов является лингвистически аннотированными, т. е. содержат различного типа разметку – морфологическую, синтаксическую, семантическую, в них могут быть выделены имена, указанные семантические роли и т. п.

NLTK обеспечивает способы доступа ко многим корпусам и распространяется с этими корпусами или их фрагментами (при использовании NLTK все корпуса по умолчанию должны храниться по следующему пути C:\Users\.....\AppData\Roaming\nltk_data\corpora).

В Приложении № 2 приведен список доступных корпусов текстов и их краткое описание.

2.7. Корпуса иноязычных текстов

NLTK включает и имеет средства работы с корпусами текстов на других языках, кроме английского. Для работы с этими корпусами нужно предварительно ознакомиться с вопросами кодирования символов в Python.

```
>>> nltk.corpus.cess_esp.words()[:10]
['El', 'grupo', 'estatal', 'Electricité_de_France', '-Fpa-', 'EDF', '-Fpt-', 'anunció', 'hoy', ',']
>>> nltk.corpus.cess_esp.words()
['El', 'grupo', 'estatal', 'Electricité_de_France', ...]
>>> nltk.corpus.indian.words('hindi.pos')
['पूरण', 'प्रतबिंध', 'हटाओ', ':', 'इराक', 'संयुक्त', ...]
>>> nltk.corpus.udhr.fileids()
['Abkhaz-Cyrillic+Abkh', 'Abkhaz-UTF8', 'Achehnese-Latin1',……]
>>> nltk.corpus.udhr.fileids()[:10]
['Abkhaz-Cyrillic+Abkh', 'Abkhaz-UTF8', 'Achehnese-Latin1', 'Achuar-Shiwiar-Latin1', 'Adja-UTF8', 'Afaan_Oromo_Oromiffa-Latin1', 'Afrikaans-Latin1', 'Aguaruna-Latin1', 'Akuapem_Twi-UTF8', 'Albanian_Shqip-Latin1',……]
>>> nltk.corpus.udhr.words('Javanese-Latin1')[11:]
['Saben', 'umat', 'manungsa', 'lair', 'kanthi', 'hak', ...]
```

Последний из рассмотренных в предыдущем примере корпусов (udhr) – это набор текстов на разных языках (300 языков) Декларации прав человека.

3. Структура корпусов текстов

Рассмотрев примеры корпусов текстов можно сделать вывод, что все они имеют разную структуру (рис. 7.3). Самый простой корпус текстов не имеет структуры, это набор текстов. Другие корпуса – это наборы текстов, разделенных по категориям языка, жанра, автора. Во многих случаях категории текстов могут пересекаться между собой, поскольку тексты могут принадлежать различным категориям. Частный случай – это когда наборы текстов распределены по временным параметрам.

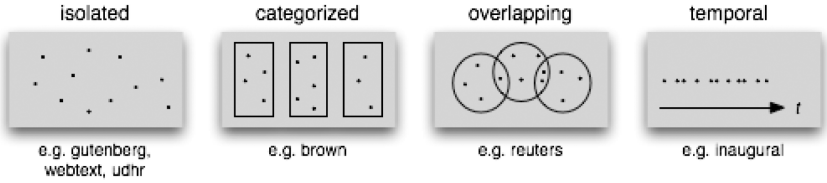

Рис. 7.3. Общие структуры корпусов текстов

3.1. Средства NLTK

Средства NLTK обеспечивают эффективные способы доступа к разным корпусам и работы с существующими и новыми корпусами. Табл. 7.2 содержит набор функций, поддерживаемых NLTK для работы с корпусами.

Таблица 7.2

Основные функции NLTK для работы с корпусами [4]

Пример использования функции	Описание
fileids()	Файлы корпуса
fileids([categories])	Файлы корпуса, соответствующие этой категории
categories()	Категории корпуса
categories([fileids])	Категории корпуса, соответствующие этим файлам

Окончание табл. 7.2

Пример использования функции	Описание
raw()	Корпус как последовательность символов
raw(fileids=[f1,f2,f3])	Последовательность символов из следующих файлов
raw(categories=[c1,c2])	Последовательность символов из следующих категорий
words()	Слова корпуса
words(fileids=[f1,f2,f3])	Слова из следующих файлов
words(categories=[c1,c2])	Слова из следующих категорий
sents()	Предложение корпуса
sents(fileids=[f1,f2,f3])	Предложение корпуса из следующих файлов
sents(categories=[c1,c2])	Предложение корпуса из следующих категорий
abspath(fileid)	Местонахождение данного файла на диске
encoding(fileid)	Кодировка (если известно)
open(fileid)	Открытие файла из корпуса для чтения
root()	Путь к месту, где установлен корпус
readme()	Содержимое файла README корпуса текстов

Различия между методами доступа к корпусам можно проиллюстрировать следующим примером.

```
>>> import nltk
>>> from nltk.corpus import*
>>> raw = gutenberg.raw(«burgess-busterbrown.txt»)
>>> raw[1:20]
'The Adventures of B'
>>> words = gutenberg.words(«burgess-busterbrown.txt»)
>>> words[1:20]
['The', 'Adventures', 'of', 'Buster', 'Bear', 'by', 'Thornton', 'W', '.', 'Burgess',
'1920', ']', 'I', 'BUSTER', 'BEAR', 'GOES', 'FISHING', 'Buster', 'Bear']
>>> sents = gutenberg.sents(«burgess-busterbrown.txt»)
>>> sents[1:20]
[['I'], ['BUSTER', 'BEAR', 'GOES', 'FISHING'], ['Buster', 'Bear',
'yawned', 'as', 'he', 'lay', 'on', 'his', 'comfortable', 'bed', 'of', 'leaves', 'and',
'watched', 'the', 'first', 'early', 'morning', 'sunbeams', 'creeping', 'through',
'the', 'Green', 'Forest', 'to', 'chase', 'out', 'the', 'Black', 'Shadows', '.'],
```

['Once', 'more', 'he', 'yawned', ',', 'and', 'slowly', 'got', 'to', 'his', 'feet', 'and', 'shook', 'himself', '.'], ['Then', 'he', 'walked', 'over', 'to', 'a', 'big', 'pine', '-', 'tree', ',', 'stood', 'up', 'on', 'his', 'hind', 'legs', ',', 'reached', 'as', 'high', 'up', 'on', 'the', 'trunk', 'of', 'the', 'tree', 'as', 'he', 'could', ',', 'and', 'scratched', 'the', 'bark', 'with', 'his', 'great', 'claws', '.'], ['After', 'that', 'he', 'yawned', 'until', 'it', 'seemed', 'as', 'if', 'his', 'jaws', 'would', 'crack', ',', 'and', 'then', 'sat', 'down', 'to', 'think', 'what', 'he', 'wanted', 'for', 'breakfast', '.'], ['While', 'he', 'sat', 'there', ',', 'trying', 'to', 'make', 'up', 'his', 'mind', 'what', 'would', 'taste', 'best', ',', 'he', 'was', 'listening', 'to', 'the', 'sounds', 'that', 'told', 'of', 'the', 'waking', 'of', 'all', 'the', 'little', 'people', 'who', 'live', 'in', 'the', 'Green', 'Forest', '.'……..]

3.2. Доступ к собственным корпусам текстов

При наличии собственного набора текстовых файлов к ним можно организовать доступ, используя вышеперечисленные методы, предварительно использовав класс NLTK PlaintextCorpusReader. Нужно знать размещения файлов на диске (в примере нужно создать папку с файлами *.htm,*html,*docпо следующему пути D:\ Jra \ Taranenko-Y \ KL \ KL). Переменной присваивается это значение (#1). Класс PlaintextCorpusReader имеет два параметра путь – к файлам и шаблон выбора файлов (# 2) и возвращает список имен файлов.

```
>>> import nltk
>>> from nltk.corpus import PlaintextCorpusReader
>>> corpus_root ='D:\Jra\Taranenko-Y\KL\KL'#1
>>> wordlists = PlaintextCorpusReader(corpus_root, '.*')#2
>>> wordlists.fileids()
["0'048.htm", '004556.html', 'about_pc-kimmo.html', 'ai00011f.htm', 'archive_article.asp.htm']
>>> wordlists.words('about_pc-kimmo.html')
['<!', 'DOCTYPE', 'HTML', 'PUBLIC', '"-//', 'W3C', ...]
```

В следующем примере показано, каким образом можно получить доступ к локальной копии корпуса PennTreebank, используя класс BracketParseCorpusReader. Для этого на диске C создайте папку compra, скопируйте в эту папку корпус treebank.

```
>>> import nltk
>>> from nltk.corpus import BracketParseCorpusReader
>>> corpus_root = r»C:\corpora\treebank»
>>> file_pattern = r».*/wsj_.*\.mrg»
>>> ptb = BracketParseCorpusReader(corpus_root, file_pattern)
>>> ptb.fileids()
['combined/wsj_0001.mrg', 'combined/wsj_0002.mrg', 'combined/
wsj_0003.mrg', 'combined/wsj_0004.mrg', 'combined/wsj_0005.mrg',
'combined/wsj_0006.mrg', 'combined/wsj_0007.mrg', 'combined/
wsj_0008.mrg', 'combined/wsj_0009.mrg', 'combined/wsj_0010.
mrg',…….]
>>> len(ptb.sents())
3914
>>> ptb.sents(fileids='combined/wsj_0001.mrg')[:19]
[['Pierre', 'Vinken', ',', '61', 'years', 'old', ',', 'will', 'join', 'the', 'board',
'as', 'a', 'nonexecutive', 'director', 'Nov.', '29', '.'], ['Mr.', 'Vinken', 'is',
'chairman', 'of', 'Elsevier', 'N.V.', ',', 'the', 'Dutch', 'publishing', 'group',
'.']]
```

4. Условное частотное распределение. Класс ConditionalFreqDist

Если тексты в корпусе разделены на различные категории (по жанру, тематике, авторами), то можно построить частотные распределения для каждой из категорий. Такие данные позволяют исследовать различия между жанрами. Условное частотное распределение – это набор частотных распределений, каждое из которых соответствует определенному «условию». Таким условием может быть категория текста.

4.1. Условия и события

Частотное распределение определяет числовые значения для каждого события (событиями можем считать употребление слов в тексте). Условное частотное распределение объединяет в пары каждое событие и условие. Вместо обработки последовательности слов (#1) обрабатываются последовательности пар (# 2).

```
>>>text = ['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', ...]        #1
>>> pairs = [('news', 'The'), ('news', 'Fulton'), ('news', 'County'), ...] #2
```

Каждая пара соответствует шаблону (condition, event). Если рассматривать корпус Brown по жанрам, то получим 15 условий (одна для каждого жанра) и 1161192 событий (одно для слова).

4.2. Подсчет слов для отдельных жанров

Используя класс ConditionalFreqDist, можно определить частоту слов для разных жанров. В случае модальных глаголов программа будет выглядеть следующим образом.

```
>>> import nltk
>>> from nltk.corpus import brown
>>> cfd = nltk.ConditionalFreqDist(
        (genre, word)
        for genre in brown.categories()
        for word in brown.words(categories=genre))
>>> genres = ['news', 'religion', 'hobbies', 'science_fiction', 'romance',
'humor']
>>> modals = ['can', 'could', 'may', 'might', 'must', 'will']
>>> cfd.tabulate(conditions=genres, samples=modals)
        can could may might must will
        news 93 86 66 38 50 389
        religion 82 59 78 12 54 71
        hobbies 268 58 131 22 83 264
science_fiction 16 49 4 12 8 16
        romance 74 193 11 51 45 43
humor 16 30 8 8 9 13
```

Тогда как для класса FreqDist () входными данными является список, то для класса ConditionalFreqDist () входными данными является список пар.

Рассмотрим отдельно только два жанра – новости и романтика. Для каждого жанра # 2 в цикле обрабатываем каждое слово этого жанра # 3 и получаем пары, содержащие жанр и слово # 1.

```
>>> genre_word = [(genre, word)                    #1
        for genre in ['news', 'romance']            #2
        for word in brown.words(categories=genre)]  #3
>>> len(genre_word)
170576
```

Пары в начале списка genre_word будут иметь форму ('news', word), если с конца списка их форма будет следующая ('romance ", word).

```
>>> genre_word[:4]
[('news', 'The'), ('news', 'Fulton'), ('news', 'County'), ('news', 'Grand')]
>>> genre_word[-4:]
[('romance', 'afraid'), ('romance', 'not'), ('romance', «"»), ('romance', '.')]
```

Можно использовать этот список пар для построения условного частотного распределения. Результаты построения сохраним в отдельной переменной cfd. Проверив значение переменной # 1, узнаем о количестве условий, а также можем пересмотреть эти условия # 2 и убедиться, что для каждого из условий построено частотное распределение # 3.

```
>>> cfd = nltk.ConditionalFreqDist(genre_word)
>>> cfd                                                    #1
<ConditionalFreqDist with 2 conditions>
>>> cfd.conditions()                                       #2
['news', 'romance']
>>> cfd['news']                                            #3
<FreqDist with 14394 samples and 100554 outcomes>
>>> cfd['romance']
<FreqDist with 8452 samples and 70022 outcomes>
>>> list(cfd['romance'])
[',', '.', 'the', 'and', 'to', 'a', 'of', '``', «"», 'was', 'I', 'in', 'he', 'had', '?', 'her',
'that', 'it', 'his', 'she', 'with', 'you', 'for', 'at', 'He', 'on', 'him', 'said', '!', '--',
'be', 'as', ';', 'have', 'but', 'not', 'would', 'She', 'The', ,…..]
>>>cfd['romance']['could']
193
```

С помощью условного частотного распределения можно исследовать употребление слов во временном промежутке. Исследуем слова America и citizen. Сначала превращаем все слова корпуса речей президентов США к одному виду # 1 и проверяем начальные буквы слов для учета различных форм American'sandCitizens. Далее строим условное частотное распределение, и результаты представляем в графическом виде (рис. 7.4).

```
>>> from nltk.corpus import inaugural
>>> cfd = nltk.ConditionalFreqDist(
```

```
        (target, fileid[:4])
        for fileid in inaugural.fileids()
        for w in inaugural.words(fileid)
        for target in ['america', 'citizen']
        if w.lower().startswith(target))              #1
>>> cfd.plot()
```

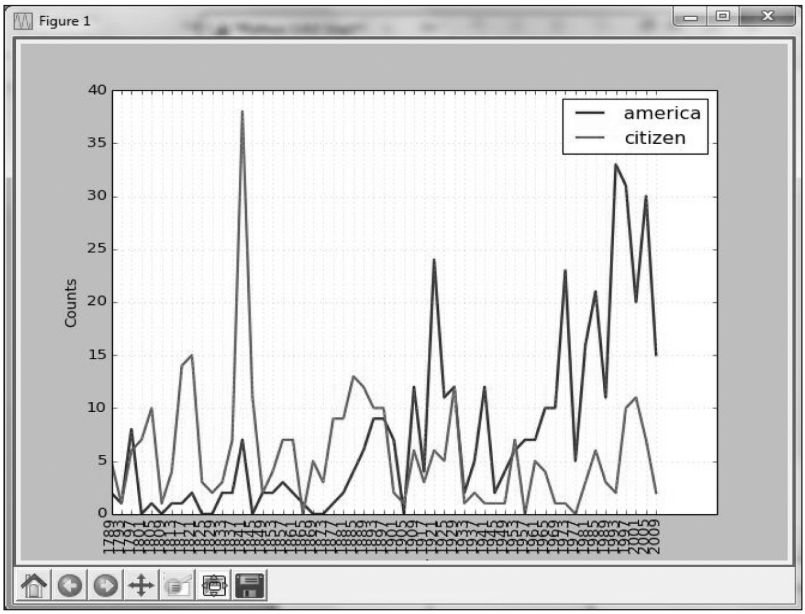

Рис. 7.4. Условное частотное распределение для определения частоты употребления слов в разные временные промежутки

Подобный график можно построить для сравнения длин слов в разных языках. Для этого также используем условное частотное распределение, но анализируем корпус udhr.

```
>>> from nltk.corpus import udhr
>>> languages = ['Chickasaw', 'English', 'German_Deutsch',
        'Greenlandic_Inuktikut', 'Hungarian_Magyar', 'Ibibio_Efik']
>>> cfd = nltk.ConditionalFreqDist(
        (lang, len(word))
        for lang in languages
        for word in udhr.words(lang + '-Latin1'))
>>> cfd.plot(cumulative=True)
```

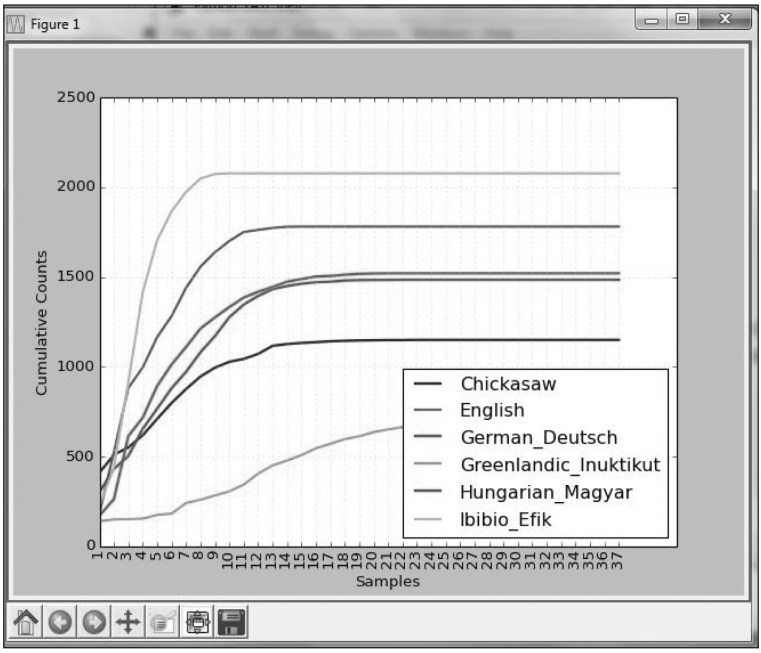

Рис. 7.5. Условное частотное распределение длин слов для разных языков

Метод plot (), а также метод tabulate () позволяют определять, какие из условий будут отображаться на экране с помощью параметра conditions = parameter. Так же определяют количество примеров для отображения с помощью параметра samples = parameter. Например, в следующей таблице отображаются длины слов до 10 символов для двух языков.

```
>>> import nltk
>>> from nltk.corpus import udhr
>>> languages = ['Chickasaw', 'English', 'German_Deutsch',
          'Greenlandic_Inuktikut', 'Hungarian_Magyar', 'Ibibio_Efik']
>>> cfd = nltk.ConditionalFreqDist(
          (lang, len(word))
          for lang in languages
          for word in udhr.words(lang + '-Latin1'))
>>> cfd.tabulate(conditions=['English', 'German_Deutsch'],
          samples=range(10), cumulative=True)
```

0 1 2 3 4 5 6 7 8 9
English 0 185 525 883 997 1166 1283 1440 1558 1638
German_Deutsch 0 171 263 614 717 894 1013 1110 1213 1275

Контрольные вопросы

1. Почему в лингвистических исследованиях для статистического анализа текстов применяют Python?
2. Как расшифровывается аббревиатура NLTK?
3. Что нужно вписать в командную строку Python для установки необходимых коллекций корпусов текстов?
4. Что показывает вкладка Collections на загрузчике коллекций корпусов текстов?
5. Из скольки приблизительно сжатых файлов содержит одна коллекция корпусов текстов и сколько дискового пространства она занимает.
6. Сколько дискового пространства занимают все коллекции корпусов текстов?
7. Структура корпуса book.
8. На что указывает частотное распределение в лингвистической статистике?
9. В какой модуль входит класс FreqDist?
10. Как в NLTK используется класс FreqDist?
11. Приведите примеры использования класса FreqDist вне корпуса.
12. Как в классе FreqDist осуществить подсчет общего количества слов в тексте?
13. Как в классе FreqDist осуществить установку списка оригинальных слов текста и сортировку их по количеству в тексте?
14. Каким оператором данные преобразуются в список?
15. Как ограничить количество слов, выводимых из списка?
16. Как определить число повторений заданного слова в тексте?
17. Как построить график частотного распределения наиболее часто используемых в тексте слов?
18. Напишите оператор для построения списка слов, которые встречаются в тексте только один раз.
19. Как выбрать из текста только те слова, длина которых (по числу знаков) больше или меньше заданной?

20. Приведите пример отбора из текста слов по двум заданным параметрам – длине слова и числе его употреблений в тексте.

21. Дайте определение коллокации. Приведите примеры.

22. Дайте определение биграммы? Приведите примеры.

23. Как преобразовать список слов в биграммы. Приведите примеры.

24. Поскольку коллокации – это частотные биграммы с учетом случаев редких слов, то какие биграммы нам нужно найти?

25. Как найти коллокации в заданном тексте?

26. Как осуществить подсчет длин слов, в тексте используя FreqDist, причем отсортировать результат в виде длин слов в порядке убывания частоты употребления слова в тексте?

27. Как осуществить подсчет длин слов в тексте, используя FreqDist, причем отсортировать результат в виде длин слов в порядке убывания частоты употребления слова в тексте с выводом длин слов и частот?

28. Как получить длину слова, употребляемого в тексте максимальное число раз.

29. Как получить относительную частоту употребления слова в тексте?

30. Дайте определение относительной частоты употребления слова в тексте.

31. Как получить доступ к корпусам текстов? Приведите пример.

32. Сколько электронных книг содержит корпус Гуттенберга?

33. Какую функцию нужно использовать для получения имен файлов (идентификаторов файлов), в которых хранятся тексты?

34. Какая функция используется для работы со словами в текстах корпуса?

35. Дайте определение термина «корконданс».

36. Как определить: среднюю длину слова; среднюю длину предложения; значение лексического разнообразия?

37. Для чего используется функция sents ()? Приведите примеры.

38. Для чего используется функция len()? Приведите примеры.

39. Как прочитать файл из корпуса? Приведите примеры.

40. Как получить доступ к корпусу текстов webtext из Интернета?

41. Как получить доступ к корпусу сообщений из чатов, созданный в NavalPostgraduateSchool?

42. Опишите возможности первого корпуса английского языка Brown.

43. Перечислите основные литературные жанры текстов корпуса английского языка Brown.

44. Как получить доступ к текстам определённого жанра корпуса Brown?

45. В чём состоит организация подсчетов употребления различных модальных глаголов для разных жанров в корпусе текстов Brown?

46. Какова структура корпуса информационного агентства Рейтер?

47. Как получить доступ к корпусу текстов агентства Рейтер?

48. Какое разделение текстов корпуса Рейтер необходимо для тренировки и тестирования алгоритмов автоматического определения тематики текста?

49. Какова структура корпуса инаугурационных приёмов президентов США?

50. Какова структура корпуса иноязычных текстов?

51. Перечислите основные функции работы с корпусами.

52. Доступ к собственным корпусам текстов.

53. Проиллюстрируйте три различных метода доступа к корпусам текстов с использованием функций raw("burgess-busterbrown.txt"), words("burgess-busterbrown.txt"), sents("burgess-busterbrown.txt").

54. Как, используя класс ConditionalFreqDist, можно определить частоту слов для разных жанров?

55. Как с помощью условного частотного распределения можно исследовать употребление слов во временном промежутке?

56. Как с помощью условного частотного распределения получить распределение длин слов для разных языков.

Практическое занятие №7

Задание: Осуществить анализ двух жанров корпуса **Brown** (новости, романтика) для определения, какие из дней недели более романтические, а содержат больше новостей. Для решения этой задачи нужно построить условное частотное распределение, где условиями являются жанры, а событиями являются дни недели. Результаты представить в табличной и графической формах.

ПРИЛОЖЕНИЕ № 1. ФУНКЦИИ КЛАССА FREQDIST

Пример	Пояснения
fdist = FreqDist(samples)	Построить частотное распределение на основе данных samples
fdist.inc(sample)	Увеличить значение для данного случая sample
fdist['monstrous']	Сколько раз встречается данный пример sample?
fdist.freq('monstrous')	Частота для данного примера sample
fdist.N()	Общее количество учтенных случаев
fdist.keys()	Примеры отсортированы по частоте по убыванию
for sample in fdist:	Перебор всех примеров по частоте по убыванию
fdist.max()	Пример с максимальным количеством
fdist.tabulate()	Представить частотное распределение в виде таблицы
fdist.plot()	Построить графическое изображение частотного распределения
fdist.plot(cumulative=True)	Построить графическое изображение частотного распределения с накоплением
fdist1 < fdist2	Проверка или примеры fdist1 встречаются с меньшей частотой, чем в fdist2

Функции построения и отображения условных частотных распределений.

Пример использования	Описание
cfdist = ConditionalFreqDist(pairs)	Создать условное частотное распределение из списка пар
cfdist.conditions()	Отсортированный список условий
cfdist[condition]	Частотное распределение для указанного условия

Пример использования	Описание
cfdist[condition][sample]	Частота для указанного примера и указанного условия
cfdist.tabulate()	Представление условного частотного распределения в виде таблицы
cfdist.tabulate(samples, conditions)	Представление условного частотного распределения в виде таблицы для указанных условий и примеров
cfdist.plot()	Построение графического представления условного частотного распределения
cfdist.plot(samples, conditions)	Построение графического представления условного частотного распределения для указанных условий и примеров
cfdist1 < cfdist2	Сравнение частот для примеров в различных условных частотных распределениях.

ПРИЛОЖЕНИЕ № 2. ПЕРЕЧЕНЬ КОРПУСОВ ТЕКСТОВ, КОТОРЫЕ РАСПРОСТРАНЯЮТСЯ ВМЕСТЕ С NLTK

Corpus	Compiler	Contents
Brown Corpus	Francis, Kucera	15 genres, 1.15M words, tagged, categorized
CESS Treebanks	CLiC-UB	1M words, tagged and parsed (Catalan, Spanish)
Chat-80 Data Files	Pereira & Warren	World Geographic Database
CMU Pronouncing Dictionary	CMU	127k entries
CoNLL 2000 Chunking Data	CoNLL	270k words, tagged and chunked
CoNLL 2002 Named Entity	CoNLL	700k words, pos- and named-entity-tagged (Dutch, Spanish)
CoNLL 2007 Dependency Treebanks (sel)	CoNLL	150k words, dependency parsed (Basque, Catalan)
Dependency Treebank	Narad	Dependency parsed version of Penn Treebank sample
Floresta Treebank	Diana Santos et al	9k sentences, tagged and parsed (Portuguese)
Gazetteer Lists	Various	Lists of cities and countries
Genesis Corpus	Misc web sources	6 texts, 200k words, 6 languages
Gutenberg (selections)	Hart, Newby, et al	18 texts, 2M words
Inaugural Address Corpus	CSpan	US Presidential Inaugural Addresses (1789-present)
Indian POS-Tagged Corpus	Kumaran et al	60k words, tagged (Bangla, Hindi, Marathi, Telugu)
MacMorpho Corpus	NILC, USP, Brazil	1M words, tagged (Brazilian Portuguese)
Movie Reviews	Pang, Lee	2k movie reviews with sentiment polarity classification

Corpus	Compiler	Contents
Names Corpus	Kantrowitz, Ross	8k male and female names
NIST 1999 Info Extr (selections)	Garofolo	63k words, newswire and named-entity SGML markup
NPS Chat Corpus	Forsyth, Martell	10k IM chat posts, POS-tagged and dialogue-act tagged
PP Attachment Corpus	Ratnaparkhi	28k prepositional phrases, tagged as noun or verb modifiers
Proposition Bank	Palmer	113k propositions, 3300 verb frames
Question Classification	Li, Roth	6k questions, categorized
Reuters Corpus	Reuters	1.3M words, 10k news documents, categorized
Roget's Thesaurus	Project Gutenberg	200k words, formatted text
RTE Textual Entailment	Dagan et al	8k sentence pairs, categorized
SEMCOR	Rus, Mihalcea	880k words, part-of-speech and sense tagged
Senseval 2 Corpus	Pedersen	600k words, part-of-speech and sense tagged
Shakespeare texts (selections)	Bosak	8 books in XML format
State of the Union Corpus	CSPAN	485k words, formatted text
Stopwords Corpus	Porter et al	2,400 stopwords for 11 languages
Swadesh Corpus	Wiktionary	comparative wordlists in 24 languages
Switchboard Corpus (selections)	LDC	36 phonecalls, transcribed, parsed
Univ Decl of Human Rights	United Nations	480k words, 300+ languages
Penn Treebank (selections)	LDC	40k words, tagged and parsed
TIMIT Corpus (selections)	NIST/LDC	audio files and transcripts for 16 speakers
VerbNet 2.1	Palmer et al	5k verbs, hierarchically organized, linked to WordNet
Wordlist Corpus	OpenOffice.org et al	960k words and 20k affixes for 8 languages
WordNet 3.0 (English)	Miller, Fellbaum	145k synonym sets

ЛИТЕРАТУРА

1. Головин Б.А. Язык и статистика / Б.А. Головин. – М.: Просвещение, 1971 г. – 190 с.

2. Б.Л. ван дер Варден. Математическая статистика / Б.Л. ван дер Варден. – М.: Изд. ин. лит. – 427 с.

3. Форсье Дж. Django. Разработка веб приложений на Python / Дж. Форсье, П. Биссекс, У. Чан; пер. с англ. – СПб.: Символ−Плюс, 2010. − 456 с.

4. Steven Bird, Ewan Klein, Edward Loper, Natural Language Processing with Python. – O'Reilly., 2009. – 502 s.

Навчальне видання

Тараненко Юрій Карлович
Тарнопольський Олег Борисович

ЛІНГВІСТИЧНА СТАТИСТИКА

Навчальний посібник

(російською мовою)

Редактор О.О. Шевцова
Комп'ютерна верстка О.М. Гришкіної

Підписано до друку 5.11.2014. Формат 60×84/16.
Ум. друк. арк. 6,51. Тираж 100 пр. Зам. № .

ПВНЗ «Дніпропетровський університет імені Альфреда Нобеля».
49000, м. Дніпропетровськ, вул. Набережна В.І. Леніна, 18.
Тел. (056) 778-58-66, e-mail: rio@duep.edu
Свідоцтво ДК № 4611 від 05.09.2013 р.

Віддруковано у ТОВ «Роял Принт».
49052, м. Дніпропетровськ, вул. В. Ларіонова, 145.
Тел. (056) 794-61-05, 04
Свідоцтво ДК № 4765 від 04.09.2014 р.