ДНЕПРОПЕТРОВСКИЙ УНИВЕРСИТЕТ
имени АЛЬФРЕДА НОБЕЛЯ

Ю.К. Тараненко
О.Б. Тарнопольский
М.О. Снятовская

ЛИНГВИСТИЧЕСКАЯ СТАТИСТИКА

• • •

СБОРНИК ЗАДАЧ

ДНЕПРОПЕТРОВСКИЙ УНИВЕРСИТЕТ
имени АЛЬФРЕДА НОБЕЛЯ

Ю.К. ТАРАНЕНКО
О.Б. ТАРНОПОЛЬСКИЙ
М.О. СНЯТОВСКАЯ

ЛИНГВИСТИЧЕСКАЯ СТАТИСТИКА

СБОРНИК ЗАДАЧ

Днепропетровск
2014

УДК 004.42
ББК 32.97-018
Т 20

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

У збірнику задач на прикладі розробленого авторами програмного пакету, написаного на мові програмування Python2.7, наведені приклади рішення і практичні завдання з усіх розділів дисципліни «Лінгвістична статистика». Задачник є хорошим посібником для вирішення практичних завдань і органічно доповнює навчальний посібник. Рішення наведених прикладів можна виконувати і в інших програмних середовищах, наприклад, в електронних таблицях MS Excel. Однак в методичному плані задачник є і практичним посібником при вивченні дисципліни «Основи програмування» спеціальності «Прикладна лінгвістика» на етапі самостійного програмування і вирішення прикладних завдань.

Тараненко Ю.К.

Т 20 Лингвистическая статистика: сборник задач / Ю.К. Тараненко, О.Б. Тарнопольский, М.О. Снятовская. – Днепропетровск: Днепропетровский университет имени Альфреда Нобеля. – 2014. – 48 с.

ISBN 978-966-434-311-1

В сборнике задач на примере разработанного авторами программного пакета, написанного на языке программирования Python2.7, приведены примеры решения и практические задания по всем разделам дисциплины «Лингвистическая статистика». Задачник является хорошим пособием для решения практических задач и органически дополняет учебное пособие. Решение приведенных примеров можно выполнять и в других программных средах, например, в электронных таблицах MS Excel. Однако в методическом плане задачник является и практическим пособием при изучении дисциплины «Основы программирования» специальности «Прикладная лингвистика» на этапе самостоятельного программирования и решения прикладных задач.

УДК 004.42
ББК 32.97-018

© Ю.К. Тараненко, О.Б. Тарнопольский, М.О. Снятовская, 2014
© Днепропетровский университет имени Альфреда Нобеля, оформление, 2014

ISBN 978-966-434-311-1

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Анализ программного обеспечения, приведенного на сайте [1] в каталоге лингвистических программ и ресурсов, показывает следующее. Среди продаваемых, и среди свободно распространяемых программ отсутствуют программы для проверки гипотезы несущественности отклонения частот выборки от среднего значения разными методами математической статистики с последующим выводом по гипотезе в целом. То же относится и к задаче определения несущественности отклонений средних значений частот в двух выборках. Соответственно, понятно, что для полной автоматизации задачи лингвистической статистики нужно связать с решением уже упомянутых задач задачу частотного анализа текста по заданным граммемам или даже тегам.

I. ИНТЕРФЕЙС ПЕРВОГО МОДУЛЯ *GRAFIKS27* ПРОГРАММНОГО КОМПЛЕКСА ЧАСТОТНОГО АНАЛИЗА СЛОВОУПОТРЕБЛЕНИЙ

Комплекс раздаётся студентам для установки на личные ПК для использовании в учебных целях. Он запрограммирован на языке Python ***версии 2.7.8. Первый модуль проекта скомпилирован в загрузочный файл Grafiks27.exe, который вместе с дополнительными файлами*** *находится в папке dist-1 и может работать без установки Python на любом компьютере, работающем под Windows начиная с Windows XP.* ***Внимание! Вместе с интерфейсной оболочкой модуля Grafiks27 по****является окно DOS, содержание которого студент должен сообщить преподавателю при неудачном старте на его ПК, чтобы получить информацию об особенностях установки на его ПК. При нормальном запуске окно DOS можно свернуть, но не закрывать.*

Интерфейс первого модуля программного комплекса **частотного анализа словоупотреблений** приведен на рис.1, он запрограммирован с учётом всех разделов пособия [2]. (Далее при ссылках на примеры и данные из пособия [2] будем употреблять только слово «пособие»).

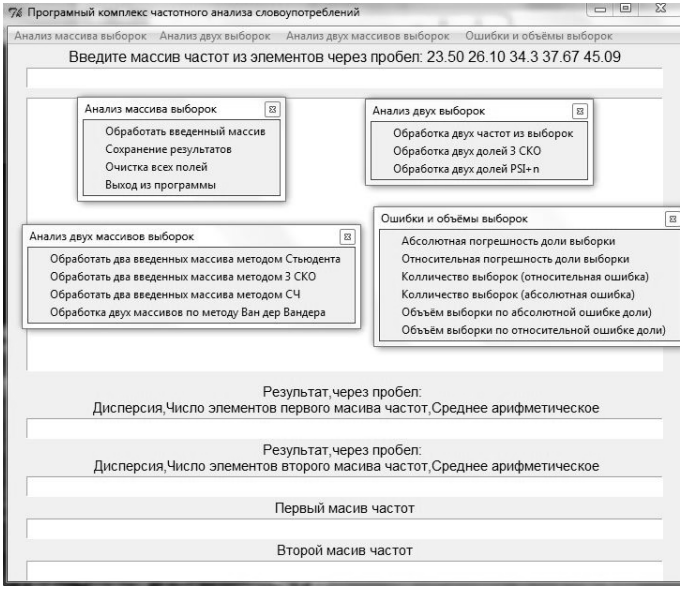

Рис. 1. Интерфейс первого модуля программного комплекса частотного анализа словоупотреблений

Основная форма содержит четыре основных меню с несколькими подменю каждое. Каждое меню через команды подменю реализует автоматизацию вычислений по методикам соответствующего раздела пособия. Так, в меню «Анализ массива выборок» реализованы методики обработки массивов, изложенные в разделе пособия «Статистическая оценка расхождений между выборочными частотами». Из этого меню можно выполнить такие общие функции: сохранения результатов расчёта, очистка всех полей формы, выход из программы. Меню «Анализ двух выборок» через соответствующие команды подменю реализует методики раздела пособия «Сравнение долей». Меню «Анализ двух массивов выборок» через соответствующие команды подменю реализует методики раздела пособия «Сравнение средних выборочных частот и частотных рядов». И наконец, меню «Ошибки и объёмы выборок» соответствуют разделу пособия «Ошибки наблюдения и определение объема выборок из текста». Все поля формы снабжены поясняющими надписями. Для ввода данных служит только первое поле.

Результат после выбора пункта меню «Обработать введенный массив» а также любого из подпунктов второго и четвертого меню выводится во втором расширенном поле. Для наглядности (рис. 1) все четыре меню приведены отдельно). При необходимости эту операцию легко осуществить простым перетаскиванием, при этом работают все восемь меню. Если текст результата не помещается в видимой части второго поля, то для его полного просмотра нужно поместить курсор во второе поле и воспользоваться колесом прокрутки «мыши».

Следует отметить оригинальную систему справки – подсказки, работающей для всех подпунктов второго и четвёртого меню, содержащего наибольшее число подменю. При возникновении затруднений в работе с указанными меню нужно очистить поле ввода и выбрать пункт меню вызывающий затруднение. При этом не стоит опасаться потери предыдущих результатов, справка-подсказка просто добавиться во второе текстовое поле к полученным ранее результатам, и после ознакомления с ними эти записи можно удалить (рис. 2).

В одном сеансе независимо от количества вычислений все результаты, пополняют второе поле. Лишнюю информацию можно удалять вручную и даже редактировать её в самом поле. Результаты можно сохранить на любом этапе расчётов, для этого нужно выбрать пункт меню «сохранение результатов» и с помощью появившегося окна диалога (рис. 3) сохранить полученный результат в нужное место на диске ПК. Без указания рас-

ширения данные сохраняются в формате txt. При необходимости можно установить расширение doc, тогда в появившемся после запуска Word файле, в окне нужно выбрать «кодированный текст» (рис. 4) и, наконец, выбрать кодировку (рис. 5).

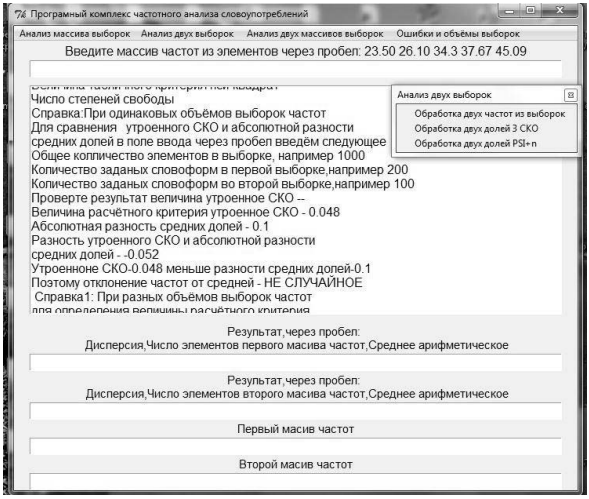

Рис. 2. Вызов справки для второго и четвёртого меню

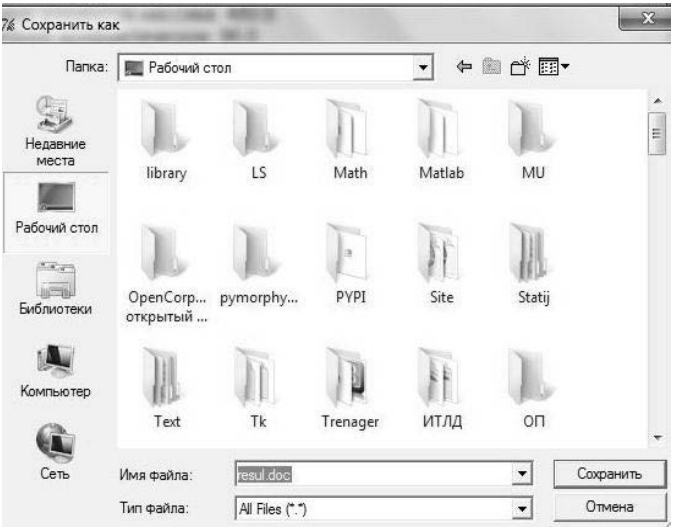

Рис. 3. Сохранение данных

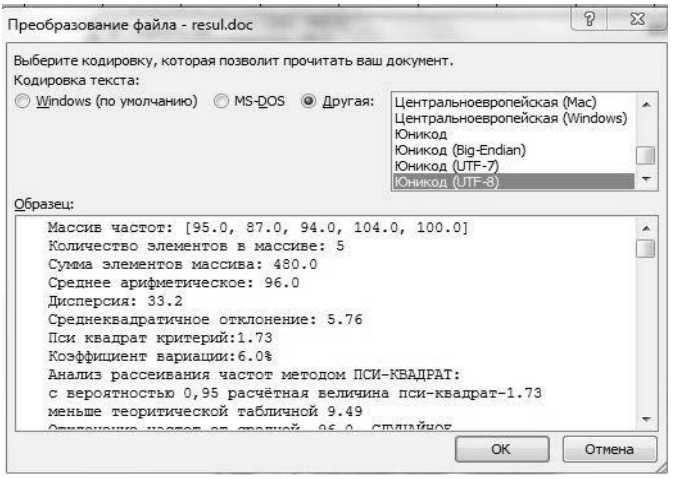

Рис. 4. Преобразование данных

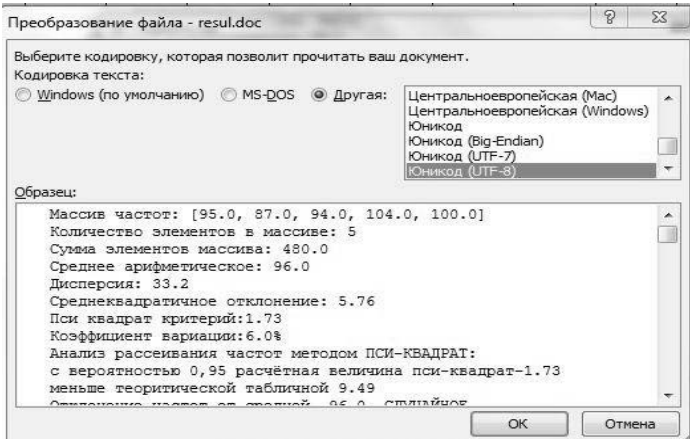

Рис. 5. Выбор кодировки

После этого в документе Word можно выбирать любой тип и размер шрифта (рис. 6).

Для получения результата, приведенного на рис. 6, нужно выполнить всего два действия – через пробел ввести массив частот: 95 87 94 104 100, выбрать подменю «Обработать введенный массив» первого меню. После чего во втором поле получим детальный статистический анализ введенного массива, фрагмент которого и приведен на рис. 6.

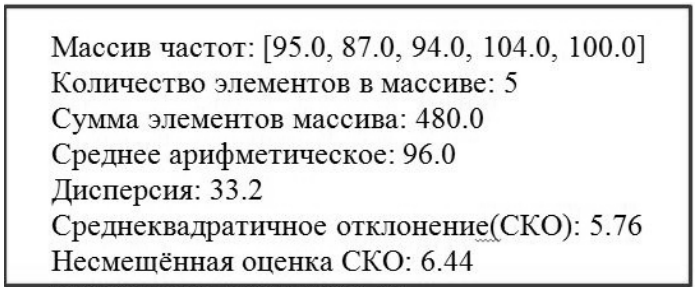

Рис. 6. Результаты расчёта одного массива

II. СТАТИСТИЧЕСКАЯ ОЦЕНКА РАСХОЖДЕНИЙ МЕЖДУ ВЫБОРОЧНЫМИ ЧАСТОТАМИ

Как видно из рис. 6, мы получили данные, аналогичные данным примера 2.1 пособия. Программа дополнительно вычисляет и другие характеристики массива, которые мы рассмотрим позже, а сейчас самостоятельно выполните задание по нахождению шести характеристик следующих массивов (табл. № 1) согласно варианту, соответствующему вашему номеру в списке.

Задание № 1. Рассчитайте для каждого массива частот: количество элементов, сумму элементов, среднюю частоту, дисперсию, среднеквадратичное отклонение, несмещённую оценку среднеквадратичного отклонения.

Таблица № 1

№	Массивы частот для расчётов			
1	71 107 95 100 70	115 102 97 95 78	81 120 117 120 72	88 95 78 106 119
2	72 107 111 95 86	78 77 103 72 119	80 107 86 80 112	102 97 107 102 92
3	100 110 98 79 91	116 81 91 87 119	108 74 90 100 90	71 82 112 90 95
4	109 76 116 115 99	82 89 100 75 119	85 120 117 114 119	89 115 93 114 89
5	97 94 76 98 94	119 99 74 85 111	78 78 118 85 102	88 85 116 100 98
6	113 70 90 118 120	82 83 91 80 83	110 108 89 112 78	84 77 105 80 104
7	72 105 108 109 76	96 116 101 120 71	74 118 118 80 85	86 79 99 99 108
8	93 111 93 101 73	78 80 82 70 118	118 100 112 102 113	78 119 75 101 113
9	102 93 93 104 86	119 89 95 95 71	107 70 118 109 95	71 115 70 97 74
10	87 76 81 90 117	102 90 116 95 109	90 114 103 100 93	79 120 120 80 80
11	95 83 79 120 104	70 80 71 83 118	73 82 81 110 79	109 106 116 106 78
12	107 111 84 74 75	119 102 119 118 118	109 101 72 109 85	85 73 80 97 71
13	95 89 112 88 110	76 107 108 89 115	117 105 81 110 77	81 93 99 92 72
14	102 80 80 106 109	88 74 117 111 118	112 71 108 80 91	81 72 111 78 107
15	91 95 108 81 82	70 95 95 112 120	107 109 70 119 87	114 99 94 92 70

Порядок выполнения задания № 1

1. На полученные четыре результата сформируйте отчёт (рис. 6) в документе Word.

2. Рассчитайте погрешность СКО относительно несмещённой оценки СКО, результаты занесите в отчёт.

3. Сравните полученные данные с данными скрипта sko.py из пособия, модернизируйте скрипт для дополнительного расчёта не смещённой оценки СКО, модернизированный скрипт занесите в отчёт.

Задание № 2. Рассчитайте для каждого массива частот показатели, приведенные в задании № 1 и дополнительно рассчи-

тайте погрешность определения средней частоты и диапазоны, в которых значение средней частоты может находиться.

Таблица № 2

№	Массивы частот для расчётов	
1	71 107 95 100 70115 102 97 95 78	81 120 117 120 7288 95 78 106 119
2	72 107 111 95 8678 77 103 72 119	80 107 86 80 112102 97 107 102 92
3	100 110 98 79 91116 81 91 87 119	108 74 90 100 9071 82 112 90 95
4	109 76 116 115 9982 89 100 75 119	85 120 117 114 11989 115 93 114 89
5	97 94 76 98 94119 99 74 85 111	78 78 118 85 10288 85 116 100 98
6	113 70 90 118 12082 83 91 80 83	110 108 89 112 7884 77 105 80 104
7	72 105 108 109 7696 116 101 120 71	74 118 118 80 8586 79 99 99 108
8	93 111 93 101 7378 80 82 70 118	118 100 112 102 11378 119 75 101 113
9	102 93 93 104 86119 89 95 95 71	107 70 118 109 9571 115 70 97 74
10	87 76 81 90 117102 90 116 95 109	90 114 103 100 9379 120 120 80 80
11	95 83 79 120 10470 80 71 83 118	73 82 81 110 79109 106 116 106 78
12	107 111 84 74 75119 102 119 118 118	109 101 72 109 8585 73 80 97 71
13	95 89 112 88 11076 107 108 89 115	117 105 81 110 7781 93 99 92 72
14	102 80 80 106 10988 74 117 111 118	112 71 108 80 9181 72 111 78 107
15	91 95 108 81 8270 95 95 112 120	107 109 70 119 87114 99 94 92 70

Порядок выполнения задания № 2

1. На полученные два результата сформируйте отчёт (рис. 5) в документе Word.

2. Определите, перекрываются ли диапазоны, в которых находятся значения частот двух исследуемых массивов, и вывод занесите в отчёт.

3. Сравните полученные данные с данными скрипта **L.py** из пособия, модернизируйте скрипт для вычисления относительной погрешности определения средней частоты, выраженной в процентах, модернизированный скрипт занесите в отчёт.

Задание № 3. По данным, приведенным в табл. № 1 (задание № 1) проверить гипотезу о том, что все четыре выборки взяты из текстов с одной и той же вероятностью, например, имен прилагательных. Иначе говоря, это проверка гипотезы о том, что все отклонения частот от их общей средней носят случайный, несущественный характер. Для проверки гипотезы применена оценка величины критерия «хи-квадрат».

Порядок выполнения задания № 3

1. На полученные четыре результата сформируйте отчёт в документе Word.

2. Изменяя данные одного из массивов, добейтесь противоположного начальному (до внесения изменений) вывода о существенности или несущественности отклонения частот в выборке от общей средней. Полученную выборку занесите в отчёт.

3. Сравните полученные Вами данные с данными скрипта **psi.py** из пособия, модернизируйте скрипт так, чтобы сразу получить результат – существенно или не существенно расхождение от средней и занесите его листинг в отчёт.

Задание № 4. По приведенным ниже данным опыта получены две частоты одного и того же явления языка в двух текстовых совокупностях, выборки из которых были равного объема. Возникает задача статистически сравнить частоты, т. е. ответить на вопрос «Существенны или случайны расхождения полученных в опыте частот?».

Таблица № 3

№ 1	№ 2	№ 3	№ 4	№ 5	№ 6	№ 7	№ 8	№ 9	№ 10
286	297	221	285	224	245	280	260	293	202
220	216	285	226	218	295	250	268	258	298
№ 11	№ 12	№ 13	№ 14	№ 15					
201	252	294	242	279					
253	207	219	224	270					

Порядок выполнения задания № 4

1. Согласно приведенному в таблице № 3 номеру в списке выберите и введите через пробел в первое поле программы (рис. 7) две частоты. Затем выберите пункт подменю «Обработка двух частот и выборок».

2. При минимальном изменении одной или обеих частот добейтесь противоположного результата (рис. 8). В предыдущем наборе первую частоту -270 нужно изменить на -240. Затем выберите пункт подменю «Обработка двух частот и выборок».

3. Сформируйте и внесите в отчёт по работе результаты расчётов, которые можно получить, выбрав в подпункте меню пункт «Сохранение результатов». Ниже приведём пример сохранения результатов в программе Word.

Для одинаковых объёмов выборок частот:

Частота первой выборки – 240.0.

Частота второй выборки – 210.0.

Величина расчётного критерия пси-квадрат – 2.0.
С вероятностью 0,95 расчётная величина пси-квадрат – 2.0 меньше теоретической табличной – 3.84.
Отклонение частот от средней – **СЛУЧАЙНОЕ.**
Для одинаковых объёмов выборок частот:
Частота первой выборки – 270.0.
Частота второй выборки – 210.0.
Величина расчётного критерия пси-квадрат – 7.5.
С вероятностью 0,95 расчётная величина пси-квадрат – 7.5 больше теоретической табличной 3.84.
Отклонение частот от средней – **НЕ СЛУЧАЙНОЕ.**

Из приведенного отчёта следует, что при уменьшении первой выборки на 30 единиц отклонение частот от средней меняется от не случайного к случайному.

Рис. 7. Обработка двух частот двух одинаковых по размеру выборок текста

4. Проверьте, совпадают ли полученные результаты с результатами работы скрипта psi12.py, результаты сравнения занесите в отчёт. Если обнаружатся расхождения, поясните причину.

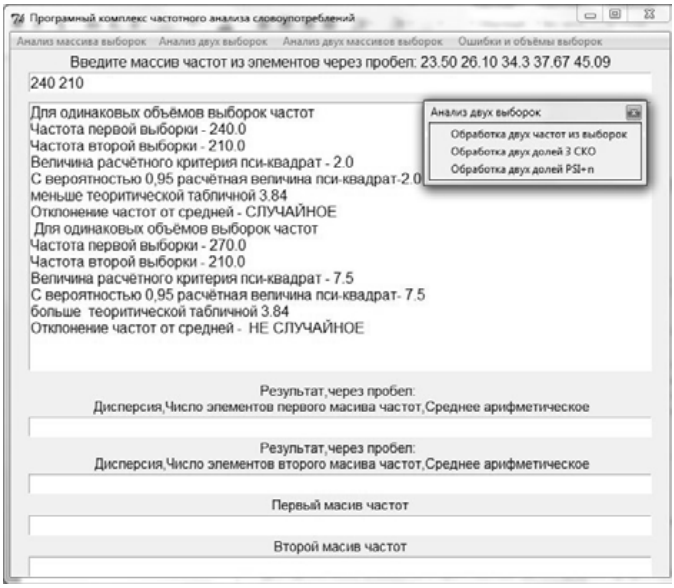

Рис. 8. Подбор большей из частот выборок, для которой выборки относятся к одному тексту

Задание № 5. В опыте получены две частоты одного и того же явления языка в двух текстовых совокупностях, выборки из которых были разного объема (выборки, разумеется, могут «отмеряться» не только количеством знаменательных слов, но иными способами, например, количеством страниц, числом строк и т. д., если страницы и строки примерно одинаковы по размеру, т. е. по числу строк в странице и по числу знаков в строке). Возникает задача статистически сравнить частоты, т. е. ответить на вопрос: Существенны или случайны расхождения полученных в опыте частот, приведённых в табл. № 3, к заданию № 4, если частоты каждой пары соответствуют парам разного объёма выборок, приведённых в следующей таблице.

Таблица № 4

№ 1	№ 2	№ 3	№ 4	№ 5	№ 6	№ 7	№ 8	№ 9	№ 10
867	719	914	568	679	720	706	759	674	775
933	839	899	628	660	526	740	721	1000	527
№ 11	№ 12	№ 13	№ 14	№ 15					
738	899	944	505	505					
576	791	615	587	987					

Порядок выполнения задания № 5

1. Для получения результата сначала внесём в программу объёмы выборок, а затем соответствующие им частоты из табл. № 4 к заданию № 4 (рис. 9). Затем выбираем пункт меню «Обработка двух частот и выборок».

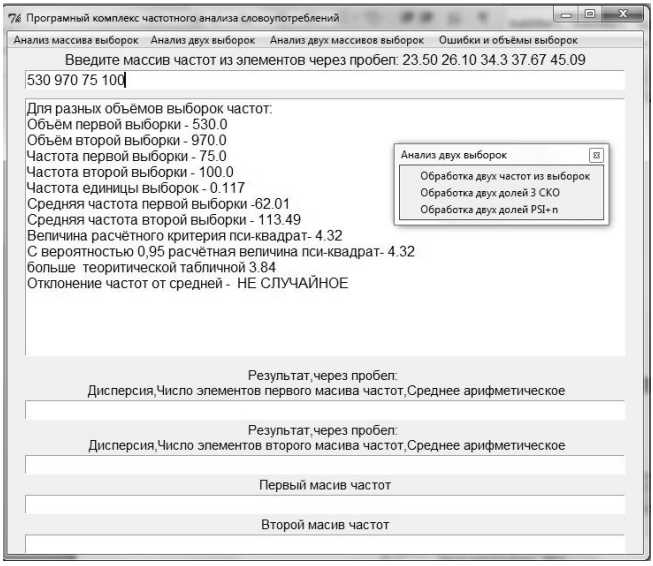

Рис. 9. Обработка двух частот с разными объёмами выборок

2. При минимальном изменении объёма одной из выборок добейтесь о результата (рис. 10), противоположного предыдущему, изменяя данные прямо в поле ввода, причем при каждом выборе пункта меню «Обработка двух частот и выборок» все предыдущие результаты сохраните вместе с вновь введёнными.

3. Полученные результаты после сохранения в документ Word переносите в отчёт по работе над заданием. Ниже приведен пример сохранения результатов в программе Word.

Для разных объёмов выборок частот:
Объём первой выборки – 580.0.
Объём первой выборки – 970.0.
Частота первой выборки – 75.0.
Частота второй выборки – 100.0.
Частота единицы выборок – 0.113.
Средняя частота первой выборки – 65.54.

Средняя частота второй выборки – 109.61.
Величина расчётного критерия пси-квадрат – 2.21.
С вероятностью 0,95 расчётная величина пси-квадрат – 2.21 меньше теоретической табличной 3.84.
Отклонение частот от средней – **СЛУЧАЙНОЕ.**
Для разных объёмов выборок частот:
Объём первой выборки – 530.0.
Объём первой выборки – 970.0.
Частота первой выборки – 75.0.
Частота второй выборки – 100.0.
Частота единицы выборок – 0.117.
Средняя частота первой выборки – 62.01.
Средняя частота второй выборки – 113.49.
Величина расчётного критерия пси-квадрат – 4.32
С вероятностью 0,95 расчётная величина пси-квадрат – 4.32 больше теоретической табличной 3.84.
Отклонение частот от средней – **НЕ СЛУЧАЙНОЕ.**

Из приведенного отчёта следует, что при увеличении объёма первой выборки на 50 не случайное отклонение от средней меняется на случайное.

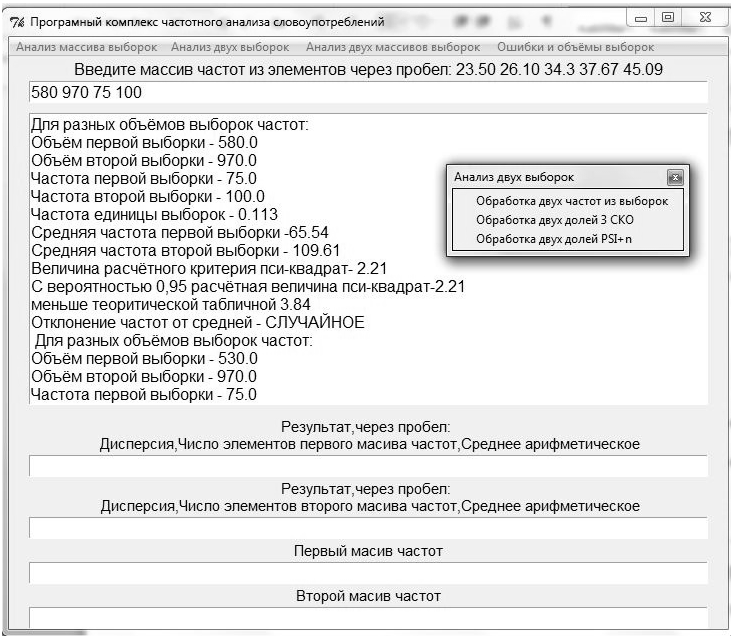

Рис. 10. Корректировка величины первой выборки прямо в поле ввода

4. Проверьте, совпадают ли полученные результаты с результатами работы скрипта psi22.py, результаты сравнения занесите в отчёт. Если обнаружатся расхождения, поясните причину.

Задание № 6. Получен ряд частот из выборок равного объема, величины которых приведены в таблице. По данному методу для оценки гипотезы о несущественности отклонения членов массива от среднего значения, нужно сравнивать коэффициент вариации членов ряда с граничным значением этого коэффициента в 40%.

Таблица № 5

№	Массивы частот для расчётов	
1	118 106 87 87 81 81 102 110 101 115	115 115 118 87 101 100 91 104 82 104
2	111 102 85 86 95 108 108 87 87 114	93 119 97 110 92 107 98 117 114 110
3	104 96 86 102 94 117 112 95 112 112	107 111 87 117 96 80 104 91 117 84
4	110 102 119 98 113 81 111 111 94 94	81 117 111 102 110 104 112 104 105 116
5	106 85 85 81 88 107 103 113 98 114	104 101 96 105 119 83 95 87 115 98
6	116 103 81 104 100 99 84 120 105 102	86 92 80 90 80 89 107 106 104 103
7	81 85 96 81 96 116 117 80 91 101	105 85 110 111 103 89 100 119 94 83
8	83 105 113 104 106 101 96 81 114 112	118 93 98 117 106 80 92 93 106 120
9	107 117 107 99 83 107 89 103 86 91	99 100 93 91 96 86 120 110 104 109
10	97 115 113 81 92 86 86 97 88 109	86 105 114 85 94 112 101 96 106 114
11	101 91 118 112 86 103 119 95 88 86	110 83 86 91 89 94 100 101 111 106
12	80 98 108 99 86 106 98 96 98 92	84 94 115 115 102 97 95 111 102 91
13	117 96 113 112 101 109 89 88 81 110	119 101 92 84 99 87 98 116 90 109
14	105 100 98 98 90 81 100 97 81 104	80 116 81 120 104 88 112 117 120 111
15	81 92 80 101 102 119 115 113 115 118	90 114 112 98 108 104 98 113 103 99

Порядок выполнения задания № 6

1. Подсчитайте значения коэффициента вариации для каждого массива с последующим определением существенного или не существенного отклонения от средней частоты.

2. Изменением параметров массива добейтесь противоположного от начального вывода о существенности отклонения от среднего значения. Дайте пояснение причинам таких изменений и занесите все результаты в отчёт.

3. Проверьте, совпадают ли полученные результаты с результатами работы скрипта var.py, результаты сравнения занесите в отчёт. Если обнаружатся расхождения, поясните причину.

III. СРАВНЕНИЕ ДОЛЕЙ

Задание № 7. Были взяты две текстовые выборки, каждая длиной в N знаменательных слов; в первой выборке оказалось n1 глаголов, во второй – n2. 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).

Исходные данные для анализа из расчёта трёх заданий на вариант приведены в табл. № 6.

Таблица № 6

№	N	n1	n2	N	n1	n2	N	n1	n2
1	1000	187	130	1000	209	150	1000	200	148
2	916	218	104	909	205	129	924	230	114
3	958	250	108	966	232	130	922	238	148
4	940	190	141	968	197	120	927	217	148
5	959	197	119	934	218	139	917	228	147
6	959	199	137	921	245	112	943	238	129
7	976	225	120	962	192	146	929	210	104
8	964	185	137	907	183	107	918	228	134
9	982	216	113	942	231	107	910	230	111
10	938	210	116	934	197	135	905	249	108
11	924	200	148	919	202	112	917	195	145
12	913	241	112	979	190	136	989	245	122
13	952	234	136	965	181	131	901	213	142
14	912	181	114	957	227	146	974	223	104
15	940	223	124	970	221	150	984	181	112

Порядок выполнения задания № 7

1. Вычислите приведенное СКО и абсолютную разность долей после вычисления путем сравнения утроенного СКО и абсолютной разности долей. Для выполнения этих действий введём в основное поле программного комплекса три числа через пробел: объём выборок (N), доля словоформы в первой выборке (n1), доля словоформы во второй выборке (n2), а затем выберите пункт меню «Обработка двух долей из выборок», как это показано на рис. 11.

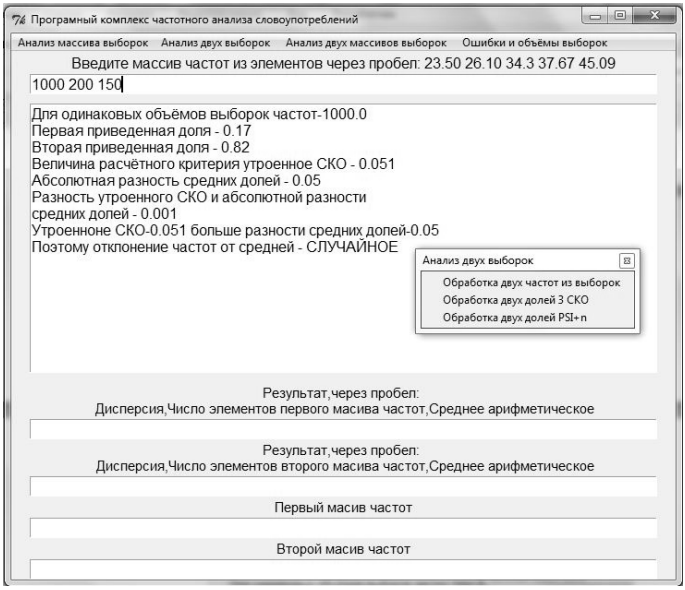

Рис. 11. Определение характера рассеивания долей около среднего

2. Изменением объёма выборок для каждого из трёх вариантов одного задания добейтесь противоположного от начального вывода о существенности отклонения от среднего значения. Дайте пояснение таких изменений и занесите все результаты в отчёт.

3. Проверьте, совпадают ли полученные результаты с результатами работы скрипта **dol.py**, результаты сравнения занесите в отчёт. Если обнаружатся расхождения, поясните причину.

Задание № 8. Провести проверку правильности выбора числа степеней свободы по методу пси-квадрат путем сравнения результатов по существенности отклонения от среднего с методом сравнения утроенного СКО с абсолютной разностью долей.

Порядок выполнения задания № 8

1. Воспользуемся табл. № 6 к задаче № 7. Сначала для своего варианта и группы условий проверяем гипотезу «о не существенности», для этого будем вводить только три параметра: общая длина выборки, количество словоформ в первой выборке, количество словоформ во второй выборке, а для расчёта будем использовать пункт меню «Обработка двух долей из выборок».

2. Далее, сохраняя первых три числа введенных, через пробел добавляем четвёртое – сначала 1 (единицу), потом 2 (двойку). Для решения воспользуемся пунктом меню «Обработка двух долей из выборок». Таким образом, для каждой из трёх групп условий (табл. № 6 к зад. № 7) получаем по три результата принятия или отказа от гипотезы о несущественности рассеивания долей около средней.

3. Сравниваем в каждой серии и трёх второй и третий результаты с первым. Результаты расчетов, распечатанные из программы, и результаты анализа необходимо занести в отчёт по работе. **Если в первом результате отклонения от среднего существенны (не случайны),** сохраняя общий объём выборки приближением меньшей доли к большей (при этом сумма не должна превышать объём общей выборки), добейтесь несущественных (случайных) отклонений от среднего и только после этого продолжайте сравнение.

IV. СРАВНЕНИЕ СРЕДНИХ ВЫБОРОЧНЫХ ЧАСТОТ И ЧАСТОТНЫХ РЯДОВ

Задание № 9. Из текстов писателя А. было взято 10 выборок по 500 знаменательных слов. Из серии текстов писателя В. было сделано столько же выборок такого же объема. Интуитивно все выборки писателя А. были определены как более или менее однородные; то же самое можно сказать о выборках из текстов писателя В.

Получены числовые данные, характеризующие частоту имен прилагательных для: писатель А., массив частот **a** и писателя В., массив частот **b** приведены в таблице. Исследователю нужно узнать, какой характер носит расхождение средних частот – случайно оно или существенно?

Таблица № 7

№	Массивы частот для расчётов	
1	84 86 84 86 77 93 94 90 74 80	79 94 92 71 73 76 90 85 72 75
2	75 94 91 71 85 80 82 89 78 86	79 76 93 91 94 73 88 80 80 90
3	72 72 70 79 90 90 94 91 94 80	85 84 85 91 78 85 90 81 74 93
4	79 73 90 77 82 76 94 79 75 89	92 86 75 78 70 72 88 79 88 84
5	93 87 83 91 72 83 94 83 76 78	83 88 95 81 74 92 83 73 90 76
6	74 89 90 87 87 82 75 74 76 80	85 89 78 75 73 72 95 89 85 80
7	78 74 89 94 86 75 90 79 92 72	79 88 73 86 92 86 78 75 73 91
8	88 91 74 94 74 83 85 94 79 92	74 74 72 81 70 87 91 91 95 80
9	78 75 86 85 71 73 81 71 80 91	85 82 87 85 77 71 90 79 88 76
10	77 84 84 78 77 71 78 84 71 72	74 94 92 90 90 91 72 87 73 82
11	80 74 74 83 88 87 75 81 91 91	89 79 74 94 70 93 81 81 75 86
12	71 79 71 71 78 92 70 70 95 83	90 73 79 78 86 80 83 81 73 73
13	90 83 71 90 87 95 85 76 94 88	84 90 71 82 92 77 78 79 87 93
14	71 91 82 72 74 86 81 75 81 85	91 76 79 73 71 87 85 83 81 91
15	77 83 82 83 75 77 72 76 85 81	79 84 72 76 86 88 76 84 90 76

Порядок выполнения задания № 9

1. Воспользуемся программным комплексом, последовательно в поле ввода введём первый массив, выберем меню «Обработать введенный массив», затем второй массив, выберем меню «Обработать введенный массив». После этого необходимо убедиться в том, что все поля основной формы комплекса заполнены результатами анализа каждого из двух массивов по отдельности и дополнительными данными для сравнительного анализа

обеих массивов в целом. Такое состояние комплекса с использованием данных из примера 5.1 пособия показано на рис. 12.

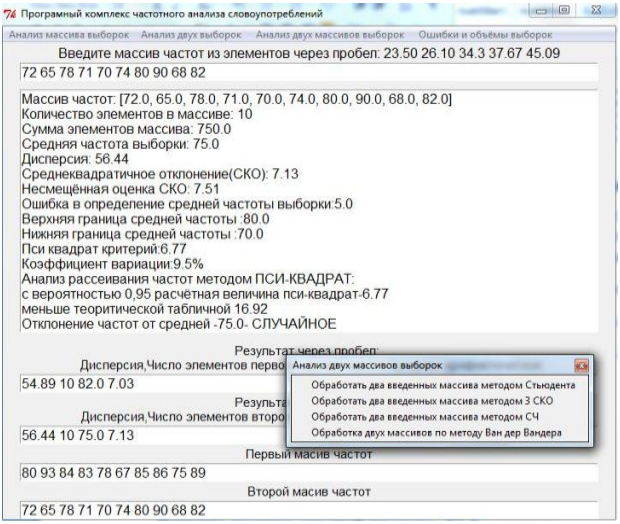

Рис. 12. Вид интерфейса комплекса перед анализом принадлежности выборок к одной генеральной совокупности

2. Прежде чем приступить к дальнейшей обработке массивов, необходимо тщательно проанализировать уже полученные по каждому массиву данные.

Необходимо детально проанализировать промежуточные данные на этапе, приведенном на рис.12. На примере данных прим. 5.1 пособия после выбора меню «Сохранение результатов» из сохранённого файла получим (пример):

Массив частот: [80.0, 93.0, 84.0, 83.0, 78.0, 67.0, 85.0, 86.0, 75.0, 89.0].

Количество элементов в массиве: 10.

Сумма элементов массива: 820.0.

Средняя частота выборки: 82.0.

Дисперсия: 54.89.

Среднеквадратичное отклонение (СКО): 7.03.

Несмещённая оценка СКО: 7.41.

Ошибка в определение средней частоты выборки: 5.0.

Верхняя граница средней частоты: 87.0.

Нижняя граница средней частоты: 77.0.

Пси-квадрат-критерий: 6.02.

Коэффициент вариации: 8.57%.

Анализ рассеивания частот методом ПСИ-КВАДРАТ:

с вероятностью 0,95 расчётная величина пси-квадрат – 6.02 меньше теоретической табличной 16.92.

Отклонение частот от средней –82.0 – СЛУЧАЙНОЕ.

Анализ рассеивания частот методом ВАРИАЦИИ:

Вариация по выборке: 8.57% меньше 40%.

Отклонение частот от средней – СЛУЧАЙНОЕ.

Массив частот: [72.0, 65.0, 78.0, 71.0, 70.0, 74.0, 80.0, 90.0, 68.0, 82.0].

Количество элементов в массиве: 10.

Сумма элементов массива: 750.0.

Средняя частота выборки: 75.0.

Дисперсия: 56.44.

Среднеквадратичное отклонение (СКО): 7.13.

Несмещённая оценка СКО: 7.51.

Ошибка в определение средней частоты выборки: 5.0.

Верхняя граница средней частоты: 80.0.

Нижняя граница средней частоты: 70.0.

Пси-квадрат-критерий: 6.77.

Коэффициент вариации: 9.5%.

Анализ рассеивания частот методом ПСИ-КВАДРАТ:

с вероятностью 0,95 расчётная величина пси-квадрат – 6.77. меньше теоретической табличной 16.92.

Отклонение частот от средней –75.0 – СЛУЧАЙНОЕ.

Анализ рассеивания частот методом ВАРИАЦИИ:

вариация по выборке: 9.5% меньше 40%.

Отклонение частот от средней – СЛУЧАЙНОЕ.

На основании анализа промежуточного отчёта можно сделать следующие выводы:

– отклонение частот от их средних значений в обоих массивах носит случайный характер. Внимание! Нельзя сравнивать массивы частот хотя бы в одном из которых, наблюдается не случайное отклонение от среднего хотя бы по одному методу. На практике исследования по такому массиву нужно повторять до тех пор, пока он не станет адекватен тексту – отклонения частот от средней будут носить случайный характер. В учебных целях данного задания изменения можно провести самостоятельно при этом, не отклоняясь от минимального и максимального значения членов массива более чем на 10–15%;

– диапазоны возможного изменения частот для обоих массивов пересекаются (77–87 и 70–80), что свидетельствует о возможной принадлежности к одной совокупности.

3. Выберем меню «Анализ двух массивов выборок» и пункт подменю «Обработать два введенных массива методом Стьюдента», получим (пример):

Метод № 1 оценки принадлежности массивов частот к одной генеральной совокупности при помощи критерия Стьюдента:

Результат сравнения выборок по методу № 1:

Для степени свободы 18, несмещённой оценки среднего квадратичного отклонения 7.86 расчётная величина критерия Стьюдента 1.99 меньше его теоретического табличного значения – 2.101.

Расхождение средних частот в двух выборках – СЛУЧАЙНОЕ.

4. Сравните полученные результаты по своему варианту с результатами, полученными при помощи скрипта **stud.py**, поясните полученные расхождения.

5. Занесите результаты в отчёт по работе.

Задание № 10. Выполнять только после завершения выполнения задания № 9 с данными к этому заданию, но с применением величины утроенного приведенного для двух массивов величины среднеквадратичного отклонения с величиной разности частот.

Порядок выполнения задания № 10

1. После обработки каждого из двух заданных массивов из табл. 9, как в задании № 9, выберите меню «Анализ двух массивов выборок» и пункт подменю «Обработать два введенных массива методом 3 СКО».

2. Провести анализ результата расчётов программным комплексом с выводами, как в примере к заданию № 9.

3. Пересчитать результат при помощи скрипта **e_p1_p2.py из** пособия, сравнить полученные результаты с программным комплексом**.** При выявлении отличий объяснить их причины. Все результаты поместить в отчёт по работе.

Задание № 11. Выполнять только после завершения выполнения задания № 9 с данными к этому заданию, но с применением анализа диапазонов средних частот СЧ.

Порядок выполнения задания № 11

1. После обработки каждого из двух заданных массивов из табл. 9, как в задании № 9, выберите меню «Анализ двух массивов выборок» и пункт подменю «Обработать два введенных массива методом СЧ».

2. Провести анализ результата расчётов программным комплексом с выводами, как в примере.

3. Провести анализ результата расчётов программным комплексом. Для этого пересчитать результат при помощи скрипта **d1_d2.py** из пособия, сравнить полученные результаты. При выявлении отличий объяснить их причины. Все результаты поместить в отчёт по работе.

Задание № 12. Выполнять только после завершения выполнения задания № 9 с данными к этому заданию, но с применение инструмента для сравнения двух частотных рядов. Этот инструмент носит название «хи-критерий» и обозначается большой греческой буквой «хи» – Х.

Порядок выполнения задания № 12

1. После обработки каждого из двух заданных массивов в отдельности, в точности, как в задании № 9, выберите меню «Анализ двух массивов выборок» и пункт подменю «Обработка двух массивов по методу Ван дер Вардена».

2. Провести анализ результата расчётов программным комплексом по всем трём методам. Сделать вывод о совпадении или различии результатов.

3. Пересчитать результат при помощи скрипта **A_B.py** из пособия**,** сравнить полученные результаты**.** При выявлении отличий объяснить их причины. Все результаты поместить в отчёт по работе.

V. ОШИБКИ НАБЛЮДЕНИЯ И ОПРЕДЕЛЕНИЕ КОЛИЧЕСТВА И ОБЪЕМА ВЫБОРОК ИЗ ТЕКСТА

Задание № 13. Было сделано по пяти выборок из двух разных текстов, каждая выборка – 500 знаменательных слов. Получены частоты имен прилагательных А, Б, приведенные в табл. № 8. Каковы ошибки наблюдения, и в каких пределах лежат действительные средние частоты?

Таблица № 8

№	Массивы частот для расчётов			
	А	В	А	В
1	58 57 56 42 46	41 24 45 41 30	56 47 40 58 50	22 22 23 24 30
2	58 56 56 46 58	43 46 30 37 32	47 44 43 44 57	30 20 29 33 27
3	51 58 43 56 44	29 41 38 47 20	50 52 52 55 47	25 35 43 26 42
4	57 59 49 49 44	40 21 30 20 45	58 45 50 44 48	38 41 38 41 48
5	48 56 53 47 43	25 24 48 43 46	59 59 41 48 48	22 45 43 20 34
6	40 43 44 54 45	31 30 22 46 31	40 53 58 58 42	43 24 40 38 24
7	59 42 59 53 46	30 36 38 25 28	50 46 44 50 41	43 32 29 44 35
8	43 50 60 49 48	44 35 35 28 38	44 60 45 46 51	43 44 35 26 38
9	59 42 57 42 55	28 31 23 37 35	54 53 44 47 50	33 45 42 48 24
10	53 49 58 57 59	22 31 49 32 41	42 59 46 44 57	39 22 45 34 48
11	48 56 54 56 40	24 25 41 39 32	54 55 47 57 42	43 33 32 48 38
12	40 54 59 48 54	25 38 33 25 27	60 47 45 55 51	47 50 32 44 41
13	42 56 52 49 50	20 36 26 46 39	43 42 41 53 50	38 40 31 24 21
14	47 41 50 58 60	30 28 35 24 27	60 45 50 45 50	49 37 46 39 21
15	40 56 52 47 51	36 25 45 22 43	46 57 48 49 44	44 45 41 31 38

Порядок выполнения задания № 13

1. В поле ввод ввести последовательно оба массива, выполняя после каждого ввода операцию обработки через подменю «Обработать введенный массив».

2. Проанализировать полученный результат, как показано в следующем примере:

Массив частот: [72.0, 65.0, 78.0, 71.0, 70.0, 74.0, 80.0, 90.0, 68.0, 82.0].

Количество элементов в массиве: 10.

Сумма элементов массива: 750.0.

Средняя частота выборки: 75.0.

Дисперсия: 56.44.

Среднеквадратичное отклонение (СКО): 7.13.

Несмещённая оценка СКО: 7.51.

Ошибка в определение средней частоты выборки: 5.1.
Верхняя граница средней частоты: 80.1.
Нижняя граница средней частоты: 69.9.
Пси-квадрат-критерий: 6.77.
Коэффициент вариации: 9.5%.
Анализ рассеивания частот методом ПСИ-КВАДРАТ:
с вероятностью 0,95 расчётная величина пси-квадрат – 6.77 меньше теоретической табличной 16.92.
Отклонение частот от средней –75.0 – **СЛУЧАЙНОЕ.**
Анализ рассеивания частот методом ВАРИАЦИИ:
Вариация по выборке: 9.5% меньше 40%.
Отклонение частот от средней – **СЛУЧАЙНОЕ.**
Массив частот: [80.0, 93.0, 84.0, 83.0, 78.0, 67.0, 85.0, 86.0, 75.0, 89.0].
Количество элементов в массиве: 10.
Сумма элементов массива: 820.0.
Средняя частота выборки: 82.0.
Дисперсия: 54.89.
Среднеквадратичное отклонение (СКО): 7.03.
Несмещённая оценка СКО: 7.41.
Ошибка в определение средней частоты выборки: 5.0.
Верхняя граница средней частоты: 87.0.
Нижняя граница средней частоты: 77.0.
Пси-квадрат-критерий: 6.02.
Коэффициент вариации: 8.57%.
Анализ рассеивания частот методом ПСИ-КВАДРАТ:
с вероятностью 0,95 расчётная величина пси-квадрат – 6.02 меньше теоретической табличной 16.92.
Отклонение частот от средней –82.0 – **СЛУЧАЙНОЕ.**
Анализ рассеивания частот методом ВАРИАЦИИ:
Вариация по выборке: 8.57% меньше 40%.
Отклонение частот от средней – **СЛУЧАЙНОЕ.**

Выделенные жирным шрифтом строки и есть решение задания, однако, это вывод – справедливый, потому что в обеих выборках отклонение от среднего случайно. В противном случае говорить о действительном среднем бессмысленно, и нужно это отметить в выводах.

3. Пересчитайте результат вручную или напишите скрипт, сравните полученные результаты. При выявлении отличий объясните их причины. Все результаты поместите в отчёт по работе.

Задание № 14. Нужно получить данные **о средней частоте глаголов в тексте** с вероятностью (надежностью) в 95% и с относительной ошибкой, не превышающей 5% ($\delta = 0,05$). Из предшествующего опыта известно, что среднее квадратичное отклонение глагола в изучаемом тексте приближенно равно σ. Сколько текстовых выборок k нужно взять, если выборочная средняя частота глагола равна $\bar{x}$?

Таблица № 9

№	№ 1	№ 2	№ 3	№ 4	№ 5	№ 6	№ 7	№ 8	№ 9
σ	13	14	17	8	15	14	8	14	9
$\bar{x}$	88	85	86	92	99	91	86	96	82
№	№ 10	№ 11	№ 12	№ 13	№ 14	№ 15			
σ	11	8	13	17	12	16			
$\bar{x}$	95	85	81	98	84	91			

Порядок выполнения задания № 14

1. В поле *ввод* ввести через пробел данные в следующей последовательности $\sigma, \delta, \bar{x}$. Затем выбираем подменю «Количество выборок (относительная ошибка)». Если вместо результатов расчётов выводится справка, следует перезапустить программный комплекс.

2. В пределах 10% уменьшать последовательно $\sigma, \delta, \bar{x}$. Определите какой из приведенных параметров имеет наибольшее влияние на количество выборок.

Задание № 15. Известно из предшествующего опыта, что доля наречий приближенно равна m **(в авторском повествовании** и описании в художественной прозе). Какую выборку нужно взять, чтобы абсолютная ошибка доли не превышала n?

Таблица № 10

№	№ 1	№ 2	№ 3	№ 4	№ 5	№ 6	№ 7	№ 8	№ 9
m	0.05	0.07	0.09	0.08	0.065	0.078	0.059	0.061	0.089
n	0.0045	0.006	0.0054	0.0045	0.0065	0.0062	0.007	0.005	0.004
№	№ 10	№ 11	№ 12	№ 13	№ 14	№ 15			
m	0.06	0.07	0.065	0.045	0.055	0.045			
n	0.005	0.006	0.007	0.008	0.004	0.005			

Порядок выполнения задания № 15

1. В поле *ввод* ввести через пробел данные из табл. 10 в следующей последовательности m, n. Затем выбираем подменю «Объём выборки по абсолютной ошибке доли». Если вместо результатов расчётов выводится справка, следует перезапустить программный комплекс.

2. В пределах 10% изменяйте последовательно m, n. Определите какой из приведенных параметров имеет наибольшее влияние на количество слов в выборке.

Задание № 16. Известно из предшествующего опыта, что доля наречий приближенно равна m (в авторском повествовании и описании в художественной прозе). Какую выборку нужно взять, чтобы относительная ошибка доли не превышала n?

Таблица № 11

№	№ 1	№ 2	№ 3	№ 4	№ 5	№ 6	№ 7	№ 8	№ 9
m	0.05	0.07	0.09	0.08	0.065	0.078	0.059	0.061	0.089
n	0.04	0.05	0.06	0.05	0.07	0.03	0.05	0.055	0.045
№	№ 10	№ 11	№ 12	№ 13	№ 14	№ 15			
m	0.06	0.07	0.065	0.045	0.055	0.045			
n	0.05	0.04	0.03	0.055	0.045	0.065			

Порядок выполнения задания № 16

1. В поле *ввод* ввести через пробел данные из табл. 10 в следующей последовательности m, n. Затем выбираем подменю «Объём выборки по абсолютной ошибке доли». Если вместо результатов расчётов выводиться справка, следует перезапустить программный комплекс.

2. В пределах 10% изменяйте последовательно m, n. Определите какой из приведенных параметров имеет наибольшее влияние на количество слов в выборке.

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

Наиболее эффективный способ исследования текстов – это использование корпусов текстов. Первый в истории представительный корпус Brown создан в 1960-е гг. Корпус состоит из 500 прозаических фрагментов в 2000 слов, взятых из текстов, опубликованных в США в 1961 г. В конце 1970-х гг. корпус был дополнен разметкой частей речи и морфологических признаков слов. Объем корпуса 1 млн словоупотреблений. Свободный доступ к корпусу возможен с сайта университета Лидс. Также свободный доступ к корпусу предоставляется через LDC (Linguistic Data Consortium) по адресу: http://wave.ldc.upenn.edu/cgi-bin/ldc/textcorpus?doc=yes&corpus=BROWN. Кроме того, Брауновский корпус распространяется на платной основе Международным компьютерным архивом современного английского языка (ICAME) (Берген, Норвегия).

Авторами данного задачника с использованием программирования Python вместе с общедоступной библиотекой, названной набором инструментов естественного языка (NLTK), разработан второй модуль *BrownNew27* программного комплекса частотного анализа словоупотреблений, основанный на возможностях, предоставляемых корпусом Brown.

VII. ИНТЕРФЕЙС ВТОРОГО МОДУЛЯ *BROWNNEW27* ПРОГРАММНОГО КОМПЛЕКСА ЧАСТОТНОГО АНАЛИЗА СЛОВОУПОТРЕБЛЕНИЙ

Комплекс раздаётся студентам для установки на личные ПК для использовании в учебных целях. Он запрограммирован на языке Python версии 2.7.8. Второй модуль проекта скомпилирован в загрузочный файл BrownNew27.exe, который вместе с дополнительными файлами находится в папке dist-2 и может работать без установки Python на любом компьютере, работающем под Windows, начиная Windows с XP. Внимание! Вместе с интерфейсной оболочкой модуля BrownNew27 появляется окно DOS, содержание которого студент должен сообщить преподавателю при неудачном старте, чтобы получить информацию об особенностях установки на его ПК. При нормальном запуске окно DOS можно свернуть, но не закрывать.

Начало работы с модулем состоит в проверке – все ли библиотеки загружены с корпуса Brown. Для этого в меню модуля выбираем пункт «Настроить соединение с корпусом» и рассматриваем информацию в окне *NLTKDownloader* (рис. 13). При необходимости библиотеки можно догрузить, используя кнопки *Download* или *Refresh*. При отсутствии Интернета библиотеки можно получить у преподавателя, создать папку по предлагаемому программой пути *DownloadDirectory*, скопировать в неё необходимые библиотеки по всем изучаемым нами корпусам. Убедившись в наличии библиотек, загружаем анализируемый текст в поле ввода текста, используя буфер обмена (ctrl+c-copy,ctrl+v-paste). После загрузки текста в поле диалогового окна программы можно его редактировать с различными целями. Например, удалить символы, подсчёт которых вести нецелесообразно или для подбора числа анализируемых слов и анализируемых символов, для удаления неинформативных участков текста и других целей, которые могут возникнуть в процессе анализа. Следует отметить два принципиально различных метода обработки текстового поля. Первый при выполнении команды «Определить жанр текста», при котором из 500 (количество устанавливается) отобранных в порядке убывания употребления слов из анализируемого текста. Из выбранных 500 выбирается только 5 (количество устанавливается), которые являются существительными – код NN (код устанавливается). Полученные таким образом слова (их может быть и меньше 5 по факту наличия) (рис. 14), обрабатываются по количеству употреблений в каждом жанре.

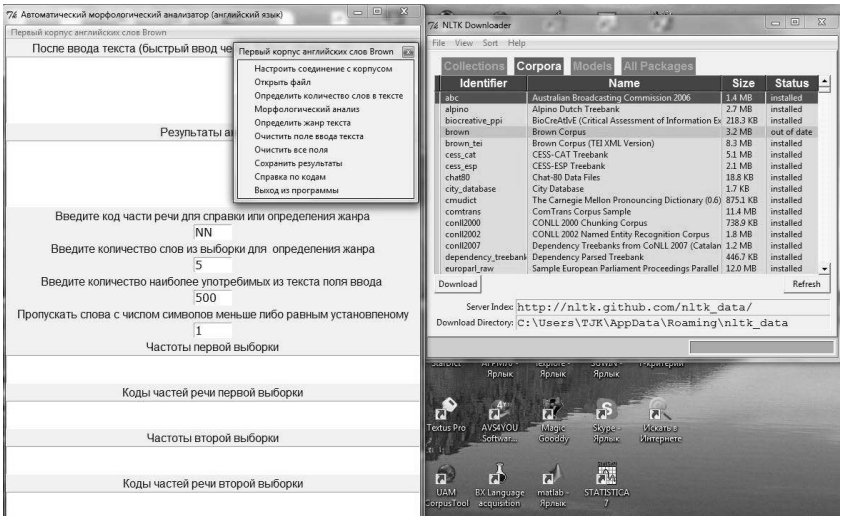

Рис. 13. Проверка загруженных библиотек корпуса Brown

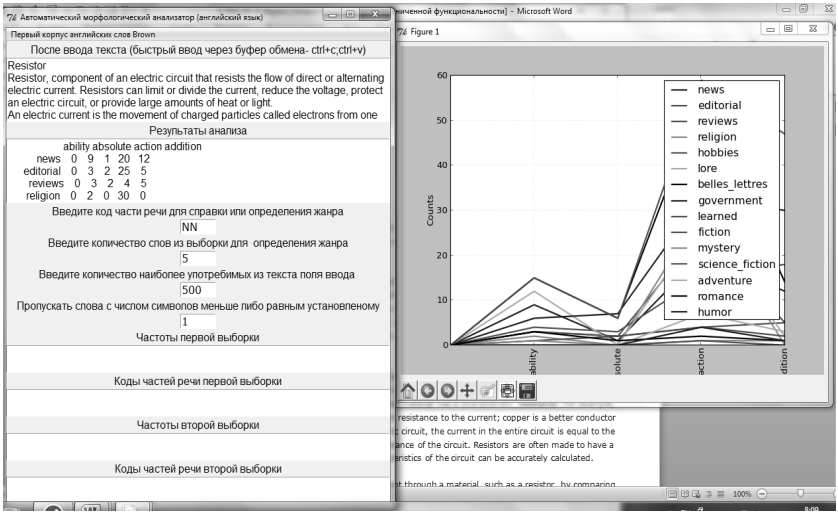

Рис. 14. Определение жанра текста с возможностью установки кода части речи и количества слов выбранной части речи для уточнения жанра

Необходимо отметить, что результаты определения жанра текста можно сохранить как в графическом (нажав на изображе-

ние дискеты в форме Figure 1) так и в текстовом вариантах (прямо скопировав текст из поля под надписью «Результаты анализа» или через меню «Сохранить результаты»).

Для осмысленного пользования программой нужно иметь табл. № 12 кодирования частей речи и символов в корпусе Brown и табл. № 13 с примерами для каждого из жанров.

Таблица № 12

Tag	Definition	Tag	Definition
.	sentence closer (. ; ? *)	NN$	possessive singular noun
(	left paren	NNS	plural noun
)	right paren	NNS$	possessive plural noun
*	not, n't	NP	proper noun or part of name phrase
--	dash	NP$	possessive proper noun
,	comma	NPS	plural proper noun
:	colon	NPS$	possessive plural proper noun
ABL	pre-qualifier (quite, rather)	NR	adverbial noun (home, today, west)
ABN	pre-quantifier (half, all)	OD	ordinal numeral (first, 2nd)
ABX	pre-quantifier (both)	PN	nominal pronoun (everybody, nothing)
AP	post-determiner (many, several, next)	PN$	possessive nominal pronoun
AT	article (a, the, no)	PP$	possessive personal pronoun (my, our)
BE	be	PP$$	second (nominal) possessive pronoun (mine, ours)
BED	were	PPL	singular reflexive/intensive personal pronoun (myself)
BEDZ	was	PPLS	plural reflexive/intensive personal pronoun (ourselves)
BEG	being	PPO	objective personal pronoun (me, him, it, them)
BEM	am	PPS	3rd. singular nominative pronoun (he, she, it, one)
BEN	been	PPSS	other nominative personal pronoun (I, we, they, you)
BER	are, art	PRP	Personal pronoun
BEZ	is	PRP$	Possessive pronoun

Окончание табл. 12

Tag	Definition	Tag	Definition
CC	coordinating conjunction (and, or)	**QL**	qualifier (very, fairly)
CD	cardinal numeral (one, two, 2, etc.)	**QLP**	post-qualifier (enough, indeed)
CS	subordinating conjunction (if, although)	**RB**	adverb
DO	do	**RBR**	comparative adverb
DOD	did	**RBT**	superlative adverb
DOZ	does	**RN**	nominal adverb (here, then, indoors)
DT	singular determiner/ quantifier (this, that)	**RP**	adverb/particle (about, off, up)
DTI	singular or plural determiner/quantifier (some, any)	**TO**	infinitive marker to
DTS	plural determiner (these, those)	**UH**	interjection, exclamation
DTX	determiner/double conjunction (either)	**VB**	verb, base form
EX	existential there	**VBD**	verb, past tense
FW	foreign word (hyphenated before regular tag)	**VBG**	verb, present participle/gerund
HV	have	**VBN**	verb, past participle
HVD	had (past tense)	**VBP**	verb, non 3rd person, singular, present
HVG	having	**VBZ**	verb, 3rd. singular present
HVN	had (past participle)	**WDT**	wh- determiner (what, which)
IN	preposition	**WP$**	possessive wh- pronoun (whose)
JJ	adjective	**WPO**	objective wh- pronoun (whom, which, that)
JJR	comparative adjective	**WPS**	nominative wh- pronoun (who, which, that)
JJS	semantically superlative adjective (chief, top)	**WQL**	wh- qualifier (how)
JJT	morphologically superlative adjective (biggest)	**WRB**	wh- adverb (how, where, when)
MD	modal auxiliary (can, should, will)	**NC**	cited word (hyphenated after regular tag)
		NN	singular or mass noun

Таблица № 13

Жанр	Описание текста
news	Chicago Tribune: *Society Reportage*
editorial	Christian Science Monitor: *Editorials*
reviews	Time Magazine: *Reviews*
religion	Underwood: *Probing the Ethics of Realtors*
hobbies	Norling: *Renting a Car in Europe*
lore	Boroff: *Jewish Teenage Culture*
belles_lettres	Reiner: *Coping with Runaway Technology*
government	US Office of Civil and Defence Mobilization: *The Family Fallout Shelter*
learned	Mosteller: *Probability with Statistical Applications*
fiction	W.E.B. Du Bois: *Worlds of Color*
mystery	Hitchens: *Footsteps in the Night*
science_fiction	Heinlein: *Stranger in a Strange Land*
adventure	Field: *Rattlesnake Ridge*
romance	Callaghan: *A Passion in Rome*
humor	Thurber: *The Future, If Any, of Comedy*

Учитывая значительный объём табл. № 12, разработчики программного комплекса снабдили модуль *BrownNew27* специальной справкой по кодам, методика работы с которой приведена на рис.15. В поле под соответствующей надписью вводится интересующий нас код, и после выполнения команды меню «Справка по кодам» в поле под надписью «Результаты анализа» получаем развёрнутую справку по коду.

Полученные справочные результаты можно сохранить в отчёте h.txt, как показано на рис. 15 выбрав пункт меню «Сохранить результаты». Текст отчёта: Справка по кодам:

NN:noun, singular, common

failure burden court fire appointment awarding compensation Mayor

interim committee fact effect airport management surveillance jail doctor intern extern night weekend duty legislation Tax Office ...

Справка по кодам:

NR: noun, singular, adverbial

Friday home Wednesday Tuesday Monday Sunday Thursday yesterday tomorrow

tonight West East Saturday west left east downtown north northeast southeast northwest North South right ...

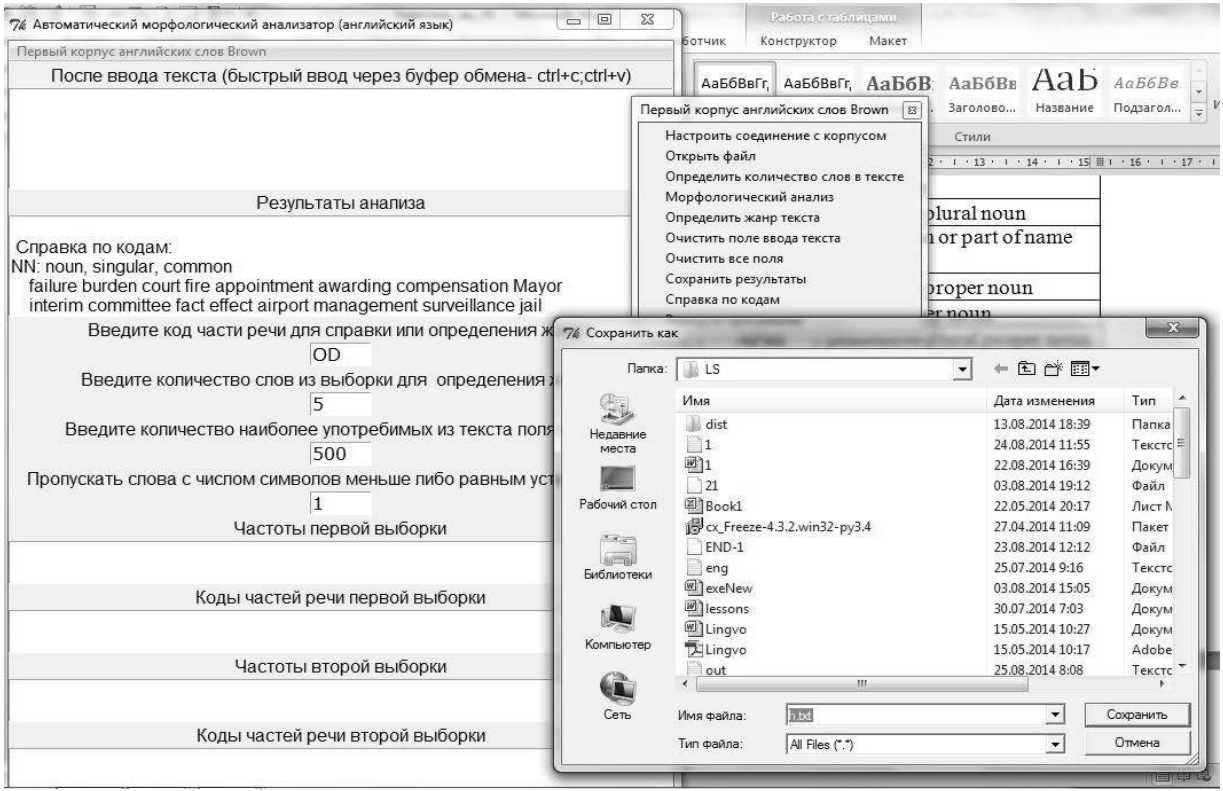

Рис. 15. Формы интерфейса для справки по коду части речи

Справка по кодам:

OD: numeral, ordinal

first 13th third nineteenth 2d 61st second sixth eighth ninth twenty-

first eleventh 50th eighteenth- Thirty-ninth 72nd 1/20th twentieth mid- 19th thousandth 350th sixteenth 701st ...

Второй метод обработки текста реализован при вызове команды меню «Морфологический анализ». Здесь возможны два варианта. Первый анализ одного текста или его фрагмента. Второй сравнительный анализ двух текстов или двух фрагментов одного текста.

Рассмотрим первый вариант анализа. Загружаем текст в поле ввода текста, используя буфер обмена (ctrl+c-copy,ctrl+v-paste). Это предпочтительный метод, поскольку при копировании из документов Word кодировка устанавливается автоматически. Возможно использование команды меню «Открыть файл», но при этом текст нужно переформатировать в формат txt, что приводит к непроизводительным затратам времени. Первый текст возьмём из энциклопедической статьи **Resistor**:

Resistor, component of an electric circuit that resists the flow of direct or alternating electric current. Resistors can limit or divide the current, reduce the voltage, protect an electric circuit, or provide large amounts of heat or light.

An electric current is the movement of charged particles called electrons from one region to another. The amount of resistance to the flow of current that a resistor causes depends on the material it is made of as well as its size and shape. Resistors are usually placed in electric circuits, which are devices formed when current moves through an electrical conductor (a material that allows the current to flow without much resistance, such as copper wire) and when the conductor makes a complete loop.

When a voltage, or electric potential, is applied to opposite ends of a circuit, it causes current to flow through the circuit. As the current flows, it encounters a certain amount of resistance from the conductor and any resistors in the circuit. Each material has a characteristic resistance. For example, wood is a bad conductor because it offers high resistance to the current; copper is a better conductor because it offers less resistance. In any electric circuit, the current in the entire circuit is equal to the voltage across that circuit divided by the resistance of the circuit. Resistors are often made to have a specific value of resistance so that the characteristics of the circuit can be accurately calculated.

Physicists sometimes explain the flow of current through a material, such as a resistor, by comparing it to water flowing through a pipe. A pressure difference maintained across two ends of the pipe by a pump is like the potential difference, or voltage, across a wire maintained by a battery. The rate of flow of water, analogous to the rate of flow of charge (current), depends on the type of pipe used. A long and thin water pipe offers more resistance than a short and thick one or a pipe that has obstructions. Similarly, the resistance of a conductor is dependent upon several factors, including its length, cross section, temperature, and a property called resistivity. Resistivity is an intrinsic characteristic of the material itself defined by the voltage divided by the density of current (current per unit cross section area) flowing across the material.

A material of high resistivity will require a higher electrical field to cause a given current density. If the resistivity of a material is known, as well as its dimensions, it can be used to calculate the resistance of a particular piece of material. The resistivity of a material is also dependent upon temperature. When a material resists the flow of current, it converts the electrical energy into other kinds of energy such as heat and light. This energy causes resistors to heat up and glow when enough current flows through them.

Resistors are designed to have a specific value of resistance. Most resistors used in electric circuits are cylindrical items a few millimeters long with wires at both ends to connect them to the circuit. Resistors are often color coded by three or four color bands that indicate the specific value of resistance. Some resistors obey Ohm's law, which states that the current density is directly proportional to the electrical field when the temperature is constant. The resistance of a material that follows Ohm's law is constant, or independent of voltage or current, and the relationship between current and voltage is linear. Modern electronic circuits depend on many devices that deviate from Ohm's law. In devices such as diodes, the current does not increase linearly with voltage and is different for two directions of current.

Resistors can help divide voltages, and when combined with other elements can help convert voltages for a specific electrical design. Resistors can also be used to provide intense light or heat. For example, the heating element in a household cooking range is a resistor, as is the tungsten filament in a common incandescent lamp. Resistors with adjustable resistance are called rheostats or potentiometers. These types of resistors are used in appliances when the current needs to be adjusted or when the resistance needs to be varied, as with lights that dim or adjustable generators.

После загрузки нужно определить объём текста через меню «Определить количество слов в тексте» (рис. 16).

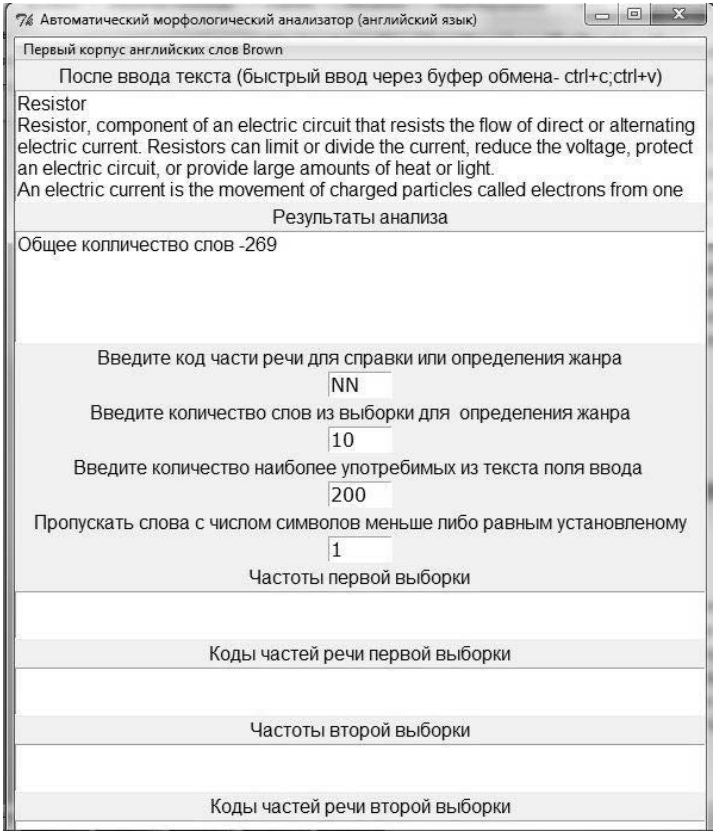

Рис. 16. Определение общего количества слов в тексте

Затем устанавливаем количество наиболее часто употребляемых слов в соотношении четыре к пяти и в соответствующем поле устанавливаем 200. Не будем учитывать слова, состоящие из одного символа и в соответствующем поле устанавливаем 1. Выбираем команду меню «Морфологический анализ», результат приведен на рис. 17 – интерфейс и на рис. 18 – распределение частот по кодам частей речи.

Анализ результатов показывает, что из 200 наиболее часто встречающихся слов в тексте примерно 75% – существительные, чтобы посмотреть какие именно слова относит машина к данным частям речи, посмотрим отчёт (меню: «Сохранить результаты»).

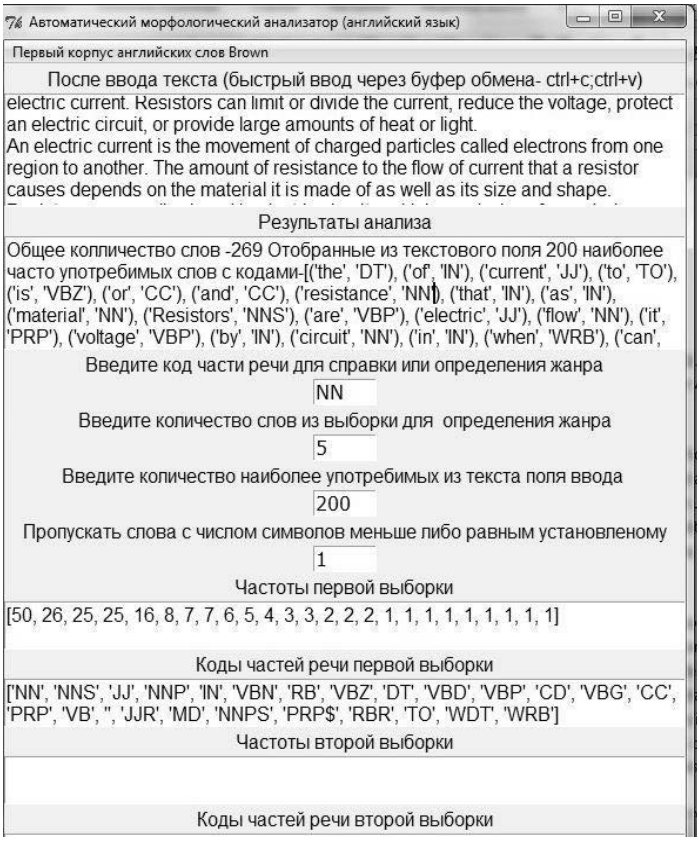

Рис. 17. Морфологический анализ энциклопедической статьи Resistor

Общее количество слов −269. Отобранные из текстового поля 200 наиболее часто употребимых слов с кодами − [('the', 'DT'), ('of', 'IN'), ('current', 'JJ'), ('to', 'TO'), ('is', 'VBZ'), ('or', 'CC'), ('and', 'CC'), ('resistance', 'NN'), ('that', 'IN'), ('as', 'IN'), ('material', 'NN'), ('Resistors', 'NNS'), ('are', 'VBP'), ('electric', 'JJ'), ('flow', 'NN'), ('it', 'PRP'), ('voltage', 'VBP'), ('by', 'IN'), ('circuit', 'NN'), ('in', 'IN'), ('when', 'WRB'), ('can', 'MD'), ('conductor', 'VB'), ('be', 'VB'), ('electrical', 'JJ'), ('resistors', 'NNS'), ('through', 'IN'), ('with', 'IN'), ('The', 'DT'), ('across', 'NNS'), ('an', 'DT'), ('circuit.', 'NNP'), ('pipe', 'NN'), ('resistance.', 'NNP'), ('specific', 'JJ'), ('such', 'JJ'), ('used', 'JJ'), ('Ohm\xe2\x80\x99s', 'NNS'), ('called', 'VBN'), ('causes', 'NNS'), ('circuits', 'NNS'), ('devices', 'NNS'), ('ends', 'NNS'), ('energy', 'NN'), ('from', 'IN'), ('heat', 'NN'), ('its', 'PRP$'),

('offers', 'NNS'), ('on', 'IN'), ('resistivity', 'NN'), ('resistor', 'NN'), ('value', 'NN'), ('water', 'NN'), ('For', 'IN'), ('In', 'NNP'), ('Resistor', 'NNP'), ('When', 'NNP'), ('adjustable', 'JJ'), ('also', 'RB'), ('amount', 'NN'), ('any', 'DT'), ('because', 'IN'), ('characteristic', 'JJ'), ('color', 'NN'), ('copper', 'NN'), ('cross', 'NN'), ('current.', 'NNP'), ('density', 'NN'), ('dependent', 'NN'), ('depends', 'NNS'), ('difference', 'NN'), ('divide', 'NN'), ('divided', 'VBD'), ('example', 'NN'), ('field', 'NN'), ('flowing', 'NN'), ('flows', 'VBZ'), ('for', 'IN'), ('has', 'VBZ'), ('have', 'VBP'), ('help', 'VBN'), ('high', 'JJ'), ('law', 'NN'), ('light.', 'NNP'), ('long', 'RB'), ('made', 'VBD'), ('maintained', 'VBN'), ('material.', 'NNP'), ('needs', 'VBZ'), ('often', 'RB'), ('one', 'CD'), ('other', 'JJ'), ('potential', 'JJ'), ('provide', 'NN'), ('rate', 'NN'), ('resists', 'NNS'), ('section', 'NN'), ('temperature', 'NN'), ('two', 'CD'), ('upon', 'IN'), ('voltages', 'NNS'), ('well', 'RB'), ('which', 'WDT'), ('wire', 'NN'), ('An', 'DT'), ('As', 'NNP'), ('Each', 'NNP'), ('If', 'NNP'), ('Modern', 'NNP'), ('Most', 'NNP'), ('Physicists', 'NNPS'), ('Resistivity', 'NNP'), ('Similarly', 'NNP'), ('Some', 'NNP'), ('These', 'NNP'), ('This', 'NNP'), ('accurately', 'RB'), ('adjusted', 'VBD'), ('allows', 'NNS'), ('alternating', 'VBG'), ('amounts', 'NNS'), ('analogous', 'JJ'), ('another.', 'NNP'), ('appliances', 'NNS'), ('applied', 'VBD'), ('area', 'NN'), ('at', 'IN'), ('bad', 'JJ'), ('bands', 'NNS'), ('battery.', 'NNP'), ('better', 'RBR'), ('between', 'IN'), ('both', 'DT'), ('calculate', 'NN'), ('calculated.', 'NNP'), ('cause', 'NN'), ('certain', 'NN'), ('characteristics', 'NNS'), ('charge', 'VBP'), ('charged', 'VBN'), ('coded', 'VBN'), ('combined', 'VBN'), ('common', 'JJ'), ('comparing', 'NN'), ('complete', 'JJ'), ('component', 'NN'), ('connect', 'NN'), ('constant', 'NN'), ('constant.', 'NNP'), ('convert', 'NN'), ('converts', 'NNS'), ('cooking', 'VBG'), ('cylindrical', 'JJ'), ('defined', 'JJ'), ('density.', 'NNP'), ('depend', 'NN'), ('design.', 'NNP'), ('designed', 'VBD'), ('deviate', 'JJ'), ('different', 'JJ'), ('dim', 'NN'), ('dimensions', 'NNS'), ('diodes', 'VBZ'), ('direct', 'JJ'), ('directions', 'NNS'), ('directly', 'RB'), ('does', 'VBZ'), ('electronic', 'JJ'), ('electrons', 'NNS'), ('element', 'NN'), ('elements', 'NNS'), ('encounters', 'NNS'), ('enough', 'RB'), ('entire', 'JJ'), ('equal', 'JJ'), ('explain', 'NN'), ('factors', 'NNS'), ('few', 'JJ'), ('filament', 'NN'), ('follows', 'VBZ'), ('formed', 'VBN'), ('four', 'CD'), ('generators', 'NNS'), ('given', 'VBN'), ('glow', 'NN'), ('heat.', 'NNP'), ('heating', 'NN'), ('higher', 'JJR'), ('household', 'NN'), ('incandescent', 'NN'), ('including', 'VBG'), ('increase', 'NN'), ('independent', 'NN'), ('indicate', 'NN'), ('intense', 'NN'), ('into', 'IN'), ('intrinsic', 'JJ'), ('items', 'NNS'), ('itself', 'PRP')]

Количество слов первой выборки – 200.

Число частот – 25.

Число кодов – 25.

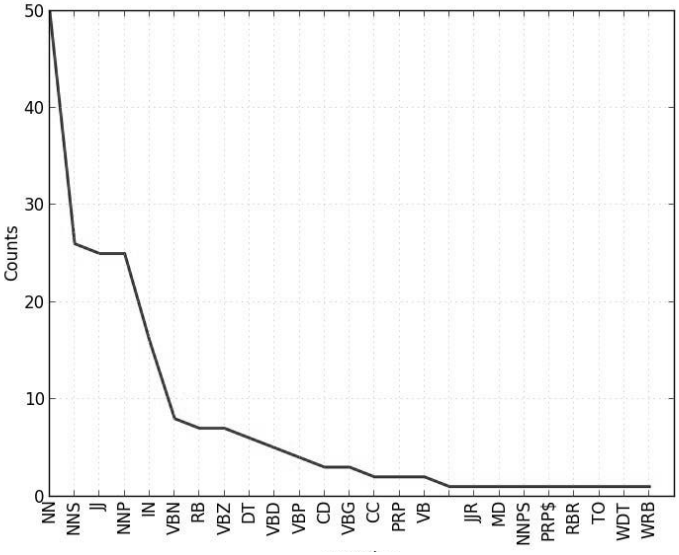

Рис. 18. Распределение частот по кодам частей речи энциклопедической статьи Resistor

Рассмотрим второй вариант анализа. Он является продолжением первого с момента, представленного на рис. 16. Выбираем функцию меню «Очистить поле ввода текста». Вводим новый текст энциклопедической статьи **Insulation** не меняя сделанные ранее установки:

Insulation

INTRODUCTION

Insulation, any material that is a poor conductor of heat or electricity, and that is used to suppress the flow of heat or electricity.

ELECTRIC INSULATION

The perfect insulator for electrical applications would be a material that is absolutely no conducting; such a material does not exist. The materials used as insulators, although they do conduct some electricity, have a resistance to the flow of electric current as much as 2.5 Ч 1024 greater than that of good electrical conductors such as silver and copper. Materials that are good conductors have a large number of free electrons (electrons not tightly bound to atoms) available to carry the current; good insulators have few such electrons. Some materials such as silicon and germanium, which have a limited number of free electrons, are semiconductors and form the basic material of transistors.

In ordinary electric wiring, plastics are commonly used as insulating sheathing for the wire itself. Very fine wire, such as that used for the winding of coils and transformers, may be insulated with a thin coat of enamel. The internal insulation of electric equipment may be made of mica or glass fibers with a plastic binder. Electronic equipment and transformers may also use a special electrical grade of paper. High-voltage power lines are insulated with units made of porcelain or other ceramic, or of glass.

The specific choice of an insulation material is usually determined by its application. Polyethylene and polystyrene are used in high-frequency applications, and mylar is used for electrical capacitors. Insulators must also be selected according to the maximum temperature they will encounter. Teflon is used in the high-temperature range of 175° to 230° C (350° to 450° F). Adverse mechanical or chemical conditions may call for other materials. Nylon has excellent abrasion resistance, and neoprene, silicone rubber, epoxy polyesters, and polyurethanes can provide protection against chemicals and moisture.

THERMAL INSULATION

Thermal insulating materials are used to reduce the flow of heat between hot and cold regions. The sheathing often placed around steam and hot-water pipes, for instance, reduces heat loss to the surroundings, and insulation placed in the walls of a refrigerator reduces heat flow into the unit and permits it to stay cold.

Thermal insulation may have to fulfill one or more of three functions: to reduce thermal conduction in the material where heat is transferred by molecular or electronic action; to reduce thermal convection currents, which can be set up in air or liquid spaces; and to reduce radiation heat transfer where thermal energy is transported by electromagnetic waves. Conduction and convection can be suppressed in a vacuum, where radiation becomes the only method of transferring heat. If the surfaces are made highly reflective, radiation can also be reduced. Thus, thin aluminum foil can be used in building walls, and reflecting metal on roofs minimizes the heating effect of the sun. Thermos bottles or Dewar flasks (*see* Cryogenics) provide insulation through an evacuated double-wall arrangement in which the walls have reflective silver or aluminum coatings. *See also* Heat Transfer.

Air offers resistance to heat flow at a rate about 15,000 times higher than that of a good thermal conductor such as silver, and about 30 times higher than that of glass. Typical insulating materials, therefore, are usually made of nonmetallic materials and are filled with small air pockets. They include magnesium carbonate, cork, felt, cotton batting, rock or glass wool, and diatomaceous earth. Asbestos

was once widely used for insulation, but it has been found to be a health hazard and has, therefore, been banned in new construction in the U.S.

In building materials, air pockets provide additional insulation in hollow glass bricks, insulating or thermopile glass (two or three sealed glass panes with a thin air space between them), and partially hollow concrete tile. Insulating properties are reduced if the air space becomes large enough to allow thermal convection or if moisture seeps in and acts as a conductor. The insulating property of dry clothing, for example, is the result of air entrapped between the fibers; this ability to insulate can be significantly reduced by moisture.

Home-heating and air-conditioning costs can be reduced by proper building insulation. In cold climates about 8 cm (about 3 in) of wall insulation and about 15 to 23 cm (about 6 to 9 in) of ceiling insulation are recommended. The effective resistance to heat flow is conventionally expressed by its R-value (resistance value), which should be about 11 for wall and 19 to 31 for ceiling insulation.

Superinsulation has been recently developed, primarily for use in space, where protection is needed against external temperatures near absolute zero. Superinsulation fabric consists of multiple sheets of aluminized Mylar, each about 0.005 cm (about 0.002 in) thick, and separated by thin spacers with about 20 to 40 layers per cm (about 50 to 100 layers per in).

Выбираем команду меню «Морфологический анализ». Получаем данные по второй выборке по тому же количеству слов (рис. 19).

Далее рассмотрим только финальную часть отчёта, которую будем использовать для дальнейшего исследования.

Общие коды для двух выборок:

'NNS', 'JJ', 'IN', 'VBN', 'RB', 'VBZ', 'DT', 'VBD', 'VBP', 'CD',

Частоты для общих кодов первой выборки:

26, 25, 16, 8, 7, 7, 6, 5, 4, 3.

Частоты для общих кодов второй выборки:

20, 21, 13, 4, 7, 2, 4, 2, 8, 19.

Частоты первой выборки:

[50, 26, 25, 25, 16, 8, 7, 7, 6, 5, 4, 3, 3, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1]

Коды первой выборки:

['NN', 'NNS', 'JJ', 'NNP', 'IN', 'VBN', 'RB', 'VBZ', 'DT', 'VBD', 'VBP', 'CD', 'VBG', 'CC', 'PRP', 'VB', '', 'JJR', 'MD', 'NNPS', 'PRP$', 'RBR', 'TO', 'WDT', 'WRB']

Частоты второй выборки:

[43, 33, 21, 20, 19, 13, 8, 7, 5, 4, 4, 4, 3, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1]

Коды второй выборки:

['NNP', 'NN', 'JJ', 'NNS', 'CD', 'IN', 'VBP', 'RB', 'VBG', 'DT', 'VB', 'VBN', 'CC', 'MD', 'PRP', 'VBD', 'VBZ', 'WDT', '', 'JJR', 'NNPS', 'PRP$', 'TO', 'WRB']

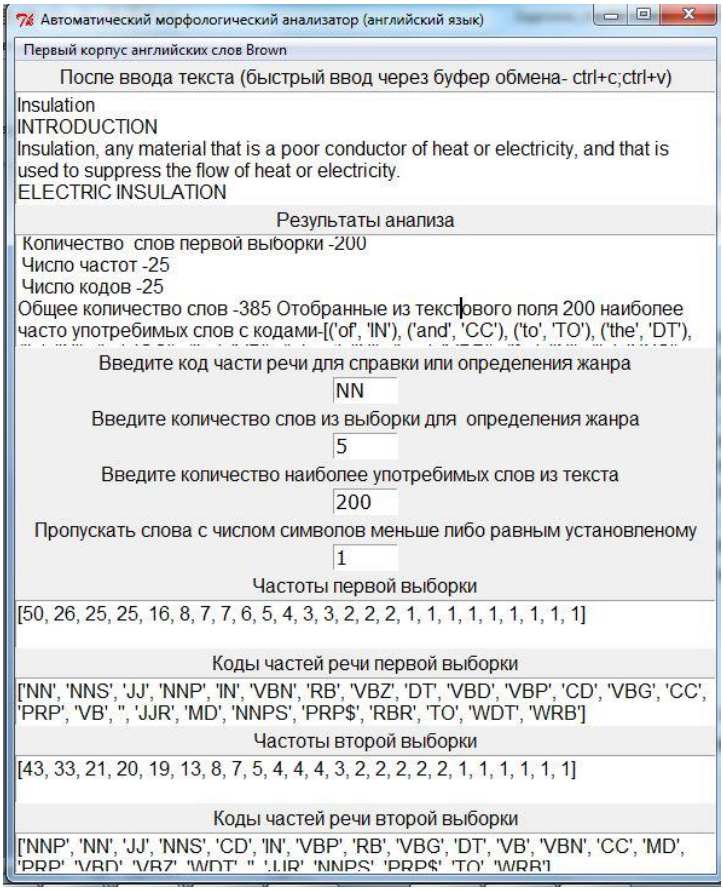

Рис. 19. Сравнительный морфологический анализ двух энциклопедических статей Resistor и Insulation

Анализ приведенного фрагмента отчёта показывает, что программа сравнила полные частотные распределения в обеих выборках и взяла наиболее значимые частоты каждой выборки с общими кодами. Теперь нужно решить вопрос, относятся ли обе статьи к одной генеральной совокупности, что может озна-

чать, что они написаны одним человеком либо обработаны одним редактором. Для этого запустим первый модуль комплекса *Grafiks27* и введем последовательно по 10 частот первой и второй выборок. Обработаем их всеми четырьмя методами (рис. 19).

Метод № 4 оценки принадлежности массивов частот к одной генеральной совокупности путём построения ранжированного ряда суммы массивов – [2.0, 2.0, 3.0, 4.0, 4.0, 4.0, 5.0, 6.0, 7.0, 7.0, 7.0, 8.0, 8.0, 13.0, 16.0, 19.0, 20.0, 21.0, 25.0, 26.0] и определения ряд позиций элементов второго массива в ранжированном ряду суммы – [0, 1, 3, 5, 10, 11, 13, 15, 16, 17] с последующим расчётом по таблице суммы весовых коэффициентов позиций второго массива – 1.09 и их сравнением с табличным значением предельной суммы весовых коэффициентов позиций второго массива –11 получим результат.

Результат сравнения выборок по методу № 4:

Расхождение средних частот в двух выборках – **СЛУЧАЙНОЕ.**

Метод № 3 оценки принадлежности массивов частот к одной генеральной совокупности путём определения погрешностей средней частоты каждого из массивов, первого – 5.8, второго – 5.1 с последующим определением диапазонов изменения средних частот для первой выборки от 4.9 до 16.5, для второй выборки от 4.9 до 15.1

Результат сравнения выборок по методу № 3:

– диапазоны погрешностей средних частот пересекаются, отклонение средних частот –**СЛУЧАЙНО.**

Метод № 2 оценки принадлежности массивов частот к одной генеральной совокупности путём сравнения утроенного среднего квадратичного отклонения 10.88 с разностью средних значений частот в выборках – 0.7.

Результат сравнения выборок по методу № 2:

Разность средних значений частот выборок 0.7 меньше утроенного среднего квадратичного отклонения 10.88, поэтому расхождение средних частот в двух выборках – **СЛУЧАЙНОЕ.**

Метод № 1 оценки принадлежности массивов частот к одной генеральной совокупности при помощи критерия Стьюдента:

Результат сравнения выборок по методу № 1:

Для степени свободы 18, несмещённой оценки среднего квадратичного отклонения 8.55 расчётная величина критерия Стьюдента 0.18 меньше его теоретического табличного значения – 2.101. Расхождение средних частот в двух выборках – **СЛУЧАЙНОЕ.**

Как следует из анализа отчёта, расхождение средних частот выборок носит случайный характер по всем четырём методам проверки, что с довольно высокой степенью вероятности может свидетельствовать о том, что автор обеих энциклопедических статей – один человек, что и соответствует действительности.

Задание № 17. Выберите два фрагмента прозы разных произведений одного автора с количеством слов, определяемым по формуле 1000-№ *10, где № – номер в списке. Провести анализ 4/5 от общего объёма текста наиболее употребляемых слов каждого фрагмента по приведенной выше методике. Провести анализ принадлежности обеих выборок к одной генеральной совокупности.

Задание 18. Выберите два фрагмента поэзии разных произведений одного автора с разным количеством слов, определ**ите частоты существительных в каждом** фрагменте. Сравните фрагменты по частотам и объёмам выборок.

Задание 19. Выберите два фрагмента поэзии разных произведений одного автора с разным количеством слов, определ**ите частоты прилагательных в каждом** фрагменте. Сравните фрагменты по частотам и объёмам выборок.

ЛИТЕРАТУРА

1. http://www.rvb.ru/soft/catalogue/catalogue.html
2. Тараненко Ю.К. Лингвистическая статистика: учебное пособие / Ю.К. Тараненко, О.Б. Тарнопольский. – Днепропетровск: Днепропетровский университет имени Альфреда Нобеля, 2014. – 112 с.

Навчальне видання

Тараненко Юрій Карлович
Тарнопольський Олег Борисович
Снятовська Мар'яна Олегівна

ЛІНГВІСТИЧНА СТАТИСТИКА

Збірник задач

(російською мовою)

Редактор О.О. Шевцова
Комп'ютерна верстка Г.М. Хомич

Підписано до друку 5.11.2014. Формат 60×84/16. Ум. друк. арк. 2,79.
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.