Институт филологии и журналистики
Саратовского государственного университета
им. Н.Г. Чернышевского
Кафедра теории, истории языка и прикладной лингвистики

Н.Б. Рогачева

Количественные методы в филологии

Учебно-методические указания

(для студентов, обучающихся по специальностям 021700 «Филология» (специализация «Русский язык и литература», «Английский язык и литература», «Немецкий язык и литература», «Французский язык и литература»), 031301 «Теоретическая и прикладная лингвистика»)

Саратов, 2011

Содержание.

Материалы для анализа.

Частотные словари.

Посмотрите на фрагмент **головы** частотного словаря и ответьте на вопросы.

Брауновский корпус	**Словарь Засориной**
The 69970	В (во) 42854
Of 36410	И 36256
And 28854	Не 19228
To 26154	На 17262
A 23363	Я 13839
An, in 21345	Быть 13307
That 10594	Что 13185
Is 10102	Он 13143
Was 9815	С (со) 12975
He 9542	А 10779

1. Какие части речи (знаменательные или служебные) в основном представлены в голове ЧС?
2. Наблюдается ли в голове ЧС совпадение частот?
3. Является ли ранг языковых единиц, входящих в голову ЧС, подвижным (т.е. если та или иная языковая единица встретится в тексте на 5 раз чаще или реже, повлияет ли это на ее ранг)?
4. Обладают ли единицы, входящие в голову ЧС стилистической окраской?

Посмотрите на фрагмент **хвоста** частотного словаря и ответьте на вопросы.

32417 1.03 прагматический adj	32432 1.03 пассивность noun	32446 1.03 безотрадный adj
32418 1.03 нервозный adj	32433 1.03 закудахтать verb	32447 1.03 пессимистический adj
32419 1.03 спагетти noun	32434 1.03 равнинный adj	32448 1.03 дворняжка noun
32420 1.03 беспроигрышный adj	32435 1.03 консерватизм noun	32449 1.03 гамбит noun
32421 1.03 неброский adj	32436 1.03 вернейший adj	32450 1.03 замочек noun
32423 1.03 радиослушатель noun	32437 1.03 негативно a dv	32451 1.03 съедение noun
32424 1.03 ровнять verb	32438 1.03 прошлепать verb	32452 1.03 порядковый adj
32425 1.03 фиброзный adj	32439 1.03 реляция noun	32453 1.03 обзываться verb
32426 1.03 апокриф noun	32440 1.03 бесчувствие noun	32454 1.03 замоpенный adj
32427 1.03 головоломный adj	32441 1.03 ненатуральный adj	32455 1.03 хмелеть verb
32428 1.03 списаться verb	32442 1.03 приобщать verb	32456 1.03 ознаменовать verb
32429 1.03 изобразиться verb	32443 1.03 реактивность noun	32457 1.03 распутник noun
32430 1.03 превалировать verb	32444 1.03 окрас noun	32458 1.03 сплетник noun
32431 1.03 исправительно adj	32445 1.03 распить verb	32459 1.03 экстремизм noun

1. Какие части речи (знаменательные или служебные) в основном представлены в хвосте ЧС?
2. Наблюдается ли в хвосте ЧС совпадение частот?
3. Является ли ранг языковых единиц, входящих в хвост ЧС, подвижным (т.е. если та или иная языковая единица встретится в тексте на 5 раз чаще или реже, повлияет ли это на ее ранг)?
4. Обладают ли единицы, входящие в хвост ЧС стилистической окраской?

Сопоставьте фрагменты головы и хвоста частотного словаря и ответьте на вопросы:

1. Увеличивается или уменьшается средняя длина слова от головы ЧС к хвосту?
2. Увеличивается или уменьшается среднее число значений у слова от головы ЧС к хвосту?

Стилеметрия.

1. Фонетика

	Драматургия	Художественная проза	Поэзия	Устная речь	Публицистика	Научная проза
Гласные	0,4248	0,4214	0,4141	0,4231	0,4167	0,4168
Согласные	0,5752	0,5786	0,5859	0,5769	0,5833	0,5832

Вопросы.

1. Какие частоты представлены в таблице – абсолютные или относительные?
2. Какие две группы стилей можно выделить по распределению гласных/согласных? Почему выделяются эти группы?

2. Морфология

2.1. Залоги

	Художественный	Научный	Публицистический
Действительный	0,42	0,47	0,55
Страдательный	0,01	0,09	0,07
Средний	0,56	0,43	0,37

Вопросы.

1. В каком стиле больше, чем в других, представлен страдательный залог? С какой стилистической доминантой это связано?

2.2. Части речи

	Художественный	Научный	Публицистический
Существительное	0,40	0,47	0,53
Прилагательное	0,15	0,23	0,22
Местоимение	0,12	0,06	0,06
Глагол	0,18	0,09	0,13
Наречие	0,07	0,08	0,04

Вопросы.

1. В каких стилях больше, чем в других, представлены существительные?
2. В каком стиле больше, чем в других, представлены глаголы?
3. Как распределение существительных и глаголов связано с отражаемой в тексте картиной мира?
4. В каком стиле больше, чем в других, представлены местоимения? С чем это связано?
5. Можно ли говорить об однозначно прямой корреляции в парах «существительное – прилагательное» и «глагол – наречие»

2.3. Падежи

	Художественный	Публицистический
Им.п.	0,28	0,18
Род.п.	0,17	0,36
Твор.п.	0,07	0,04
Предлож.п.	0,08	0,10

Вопросы.

1. В каком стиле больше, чем в другом, представлен родительный падеж? С чем это связано?

2.4. Время

	Наст.	Прош.	Буд.
Науч. (письм.)	77,8	16,5	5,7
Худож. (письм.)	6,2	91,4	2,1
Разг. (уст.)	40	47,2	12,8
Личные письма	39	46,7	14,1

Вопросы.

1. В каком стиле преобладает настоящее время? С чем это связано?
2. В каком стиле преобладает прошедшее время? С чем это связано?
3. В каком стиле представлено наибольшее разнообразие временных форм? С чем это связано?
4. Личные письма оказываются ближе к другим письменным текстам или к устным текстам разговорного стиля? Сделайте вывод: при отборе языковых средств определяющим является форма речи (письменная или устная) или функциональный стиль?

3. Синтаксис

3.1. Доли предложений разной модальности в разных стилях

	Драматургия	Поэзия	Худож. проза	Публицистика	Научный	Деловой
Пов. утвердит.	58,4	83,6	90,9	89,6	89,5	91,8
Воп. утвердит.	14,2	3,4	1,1	2,1	1,5	0,2
Побудит. утвердит.	6,6	3,6	0,6	0,9	0,08	0,02

Вопросы.

1. В каком стиле представлено наибольшее разнообразие модальностей? С чем это связано?
2. В каком стиле представлено наименьшее разнообразие модальностей? С чем это связано?
3. Какой стиль находится между «личностно-ориентированными» и «обезличенными» стилями? С какими доминантами этого стиля связано его промежуточное положение?

3.2. Доли предложений разного состава в разных стилях

	Драматургия	Поэзия	Худож. проза	Публицистика	Научный	Деловой
Двусост.полн.	45	67,1	77	78,7	74,2	80,8
Двусост.неполн.	19,7	9,7	7,2	5,8	9,8	9,6
Определенно-личн. полн.	13,6	6,8	1,6	2	1,4	1,8
Безличное полн.	7	5,9	7,4	7,9	9,6	2,2
Назывное полн.	1,6	5,1	1,5	0,9	0,9	1
Инфинитивное полн.	1,8	-	1,8	1,7	1,2	4,4
Неопред.-личн. полн.	2,4	3,2	2,3	2,2	1,2	0,2

Вопросы.

1. В каком стиле представлено наибольшее разнообразие предложений по составу? С чем это связано?
2. В каком стиле представлено наименьшее разнообразие предложений по составу? С чем это связано?
3. Можно ли на основе близости долей двусоставных неполных предложений в поэзии, научном и деловом стилях (9,7; 9,8; 9,6 соответственно) говорить о наличии у этих стилей какой-то общей стилевой доминанты? Если да, что это за

доминанта? Если нет, какими причинами можно объяснить близость количественных показателей?

4. В каком стиле больше, чем в других, представлены назывные предложения? С чем это связано?
5. В каком стиле больше, чем в других, представлены безличные предложения? С чем это связано?
6. В каком стиле больше, чем в других, представлены инфинитивные предложения? С чем это связано?

3.3. Доли позиций зависимости слов в разных стилях

	Драматургия	Поэзия	Худож. проза	Публицистика	Научный	Деловой
Вправо, 1-я	34,1	28,3	33	29,8	25,7	23,2
Вправо, 2-я	5	10,4	14	18,1	14,2	18,7
Вправо, 3-я	1	3,7	4,6	7,2	5,3	11,6
Вправо, 4-я	0,2	1,7	1,4	2,4	1,8	6,6
Вправо, 5-я и далее	0,06	1,8	0,6	1,2	0,8	5,9
Влево, 1-я	49	32,9	25,9	22,8	24,1	8,7
Влево, 2-я	9,2	12,4	14,6	11,5	17,4	9,3
Влево, 3-я	1,3	4,5	4,3	4,6	7	7,4
Влево, 4-я	0,2	1,9	1,2	1,7	2,4	3,8
Влево, 5-я и далее	0,1	2,3	0,5	0,8	1,3	4,7

Вопросы.

1. В каком стиле представлено наибольшее разнообразие позиций зависимости? С чем это связано?
2. В каком стиле представлено наименьшее разнообразие позиций зависимости? С чем это связано?
3. В каком стиле больше, чем в других, представлены «дальние» позиции зависимости? С чем это связано?

Лингвистическая типология.

Ниже представлен фрагмент таблицы с указанием степени близости между индоевропейскими языками, рассчитанными по методике А.Я. Шайкевича. Просмотрите данные в таблице и ответьте на вопросы.

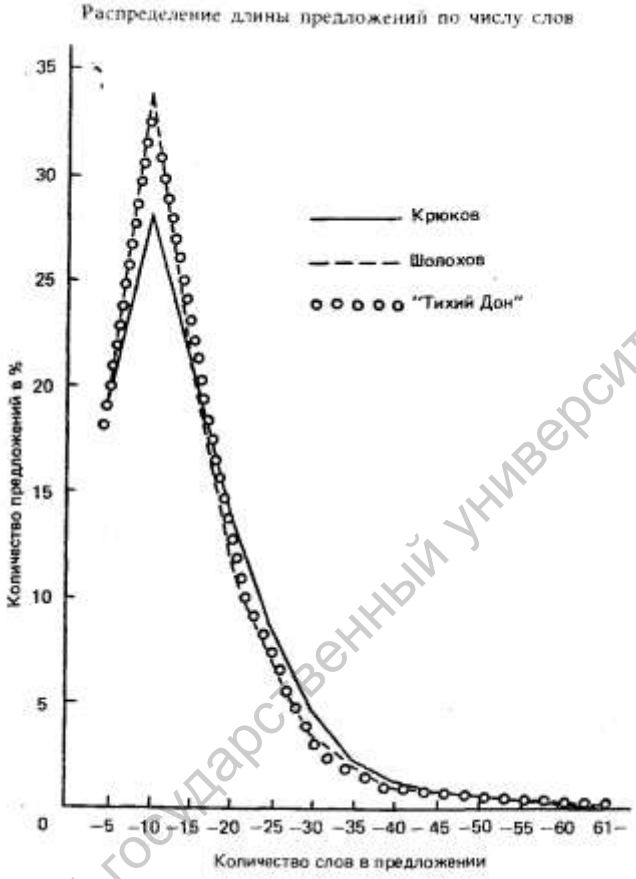
	новогреч.	фр.	исп.	дат.	швед.	нем.	польск.	рус.	баск.	англ.
новогреч.		20	19	18	18	16	14	13	-	17
фр.	17		56	22	22	24	20	19	-	21
исп.	16	62		21	21	22	19	19	-	21
дат.	13	20	19		72	56	22	22	-	49
швед.	13	20	19	85		56	22	22	-	49
нем.	13	23	21	61	61		29	22	-	48
польск.	12	15	15	18	18	19		69	-	16
рус.	11	14	14	16	16	16	64		-	14
баск.	1	4	6	2	2	2	1	1		2
англ.	13	30	27	50	53	54	23	23	-	

1. Какой язык выделяется при пороге S≥10? Чем можно объяснить его обособление?
2. Какой язык выделяется при пороге S>20? Чем можно объяснить его обособление?
3. Какие группы языков выделяются при пороге S≥40?
4. Какие подгруппы выделяются при пороге S≥70?
5. Как соотносятся группы языков, выделенные по методике А.Я. Шайкевича с традиционной генеалогической классификацией?

Атрибуция.

1. Длина предложения.

Текст	Число предл	Число слов	Сред длина предл	Макс длина предл	Макс вероят длина предл	Вероят-ть, %	Макс вероят интервал	Вероят-ть, %
К1	1781	24913	13,988	66	7	6,00	6-10	26,78
К2	1955	24352	12,456	60	6	6,45		29,11
Крюков	3736	49265	13,187	66	6	6,08		28,00
Ш1	2825	32957	11,666	95	7	8,04	6-10	34,30
Ш2	1358	18673	13,750	86	7	7,44		32,33
Шолохов	4183	51630	12,343	95	7	7,84		33,66
ТД1	1177	12720	10,807	60	6	8,92	6-10	36,62
ТД2	1168	15179	12,995	96	10	7,02		32,96
ТД4	1415	20195	14,272	102	7	6,15		27,99
ТД	3760	48094	12,790	102	6	7,07		32,23



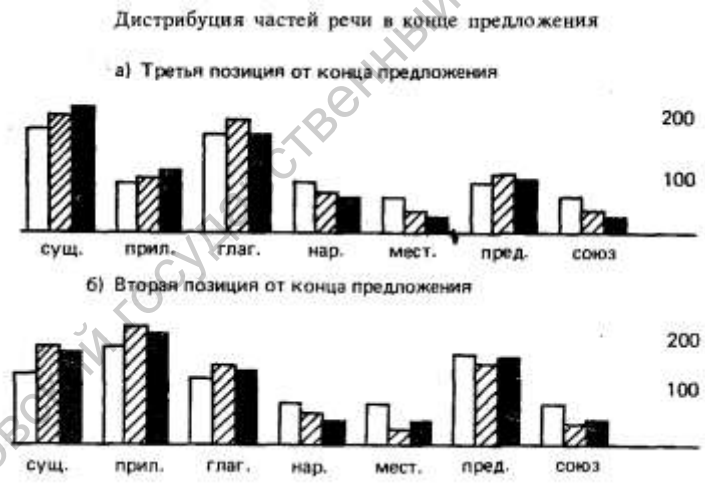
Вопросы.

1. (по таблице) Как изменяется со временем длина предложения у каждого из авторов: уменьшается/увеличивается/остается неизменной?
2. (по таблице) В текстах какого автора наблюдается наибольшее разнообразие длин предложений? тексты каких авторов в этом отношении более однообразны?
3. (по графику) Какая кривая (Крюкова или Шолохова) больше отклонилась от кривой автора «Тихого Дона»?
4. Сделайте вывод: чей стиль (Крюкова или Шолохова) ближе к стилю автора «Тихого Дона»?

2. Позиция частей речи в предложении.

Посмотрите на распределение частей речи в начале предложения в текстах трех сравниваемых авторов. Ответьте на вопросы:

1. Какой писатель (Крюков или Шолохов) по данному параметру ближе к автору «Тихого Дона»? Есть ли исключения?
2. Какое начало предложения характерно для Крюкова, но совсем не характерно для Шолохова и автора «Тихого Дона»?
3. Какое начало предложения характерно для Шолохова и автора «Тихого Дона», но совсем не характерно для Крюкова?



Посмотрите на распределение частей речи в конце предложения в текстах трех сравниваемых авторов. Ответьте на вопросы: Какой писатель (Крюков или Шолохов) по данному параметру ближе к автору «Тихого Дона»? Есть ли исключения?

Сделайте общий вывод: чей стиль (Крюкова или Шолохова) ближе к стилю автора «Тихого Дона»?

3. Соотношение частей речи.

Дистрибуция частей речи. Средние значения

[Части речи]	Крюков	Шолохов	«Тихий Дон»
Сущ[ествитель-ное]	147,85	183,60	184,50
Прил[агатель-ное]	64,20	63,75	72,25
Глаг[ол]	86,75	91,70	91,05
Нар[ечие]	51,75	38,60	34,30
Мест[оимение]	47,50	24,00	25,85
Пред[лог]	60,80	71,70	69,30
Союз	41,15	26,65	22,75
[Итого]	500,00	500,00	500,00

Изучите представленные данные и сделайте вывод: чей стиль (Крюкова или Шолохова) ближе к стилю автора «Тихого Дона»? в частотности каких частей речи различие особенно велико?

4. Сочетания частей речи.

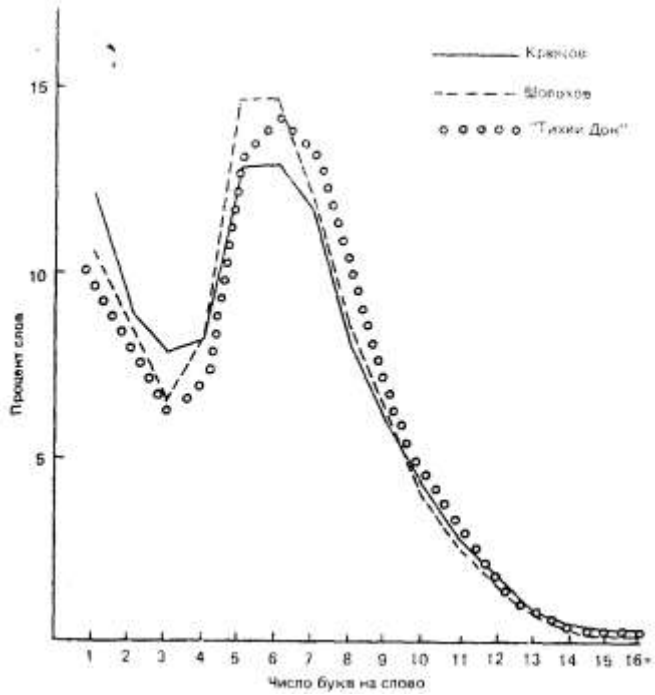
Наиболее часто встречающиеся сочетания частей речи в начале предложений

[№ п/п]	Крюков		Шолохов		«Тихий Дон»	
1	сущ/п — глаг.	117	пред.—сущ.	163	пред.— сущ.	136
2	мест.— глаг.	86	сущ/п — глаг.	121	сущ/п — глаг.	122
3	пред.— сущ.	69	пред.— прил.	49	пред.— прил.	48
4	прил.— сущ/п	44	сущ/п — нар.	36	мест.— глаг.	47
5	нар.— глаг.	43	нар.— глаг.	35	сущ/п — (,)	47
6	сущ/п — нар.	38	глаг.— сущ/п	35	прил.— сущ/п	42
7	мест.— нар.	36	сущ/п — (,)	33	пред.— мест.	40
8	глаг.— сущ/п	31	глаг.— пред.	33	глаг.— пред.	38
9	союз — нар.	27	сущ/п — пред.	32	сущ/п — сущ/п	37
10	пред.— мест.	24	сущ/п — сущ.	28	сущ/п — нар.	35
11	сущ/п — сущ/п	23	сущ/п — сущ/п	27	глаг.— сущ.	29
12	глаг.— пред.	23	глаг.— сущ.	26	нар.— глаг.	28
13	союз — сущ/п	22	мест.— глаг.	25	сущ/п — пред.	25
14	глаг.— нар.	22	прил.— сущ/п	22	сущ.— глаг.	20
15	сущ/п — сущ.	21	пред.— мест.	22	прил.— (,)	18
[Итого]		62,6%		68,7%		71,3%

Вопросы.

1. Совпадают ли самые частотные сочетания частей речи в начале предложения у всех трех авторов?
2. Запятая в самом начале предложения – явление редкое и, следовательно, показательное. Есть ли сочетания с запятой у сравниваемых авторов?
3. Есть ли у Крюкова сочетания, не входящие в 15 самых частотных в «Тихом Доне» (если да, то какие)? Есть ли такие сочетания у Шолохова (если да, то какие)?
4. Союз в самом начале предложения – явление редкое и, следовательно, показательное. Есть ли у сравниваемых авторов сочетания с союзом?
5. В текстах какого автора представлено большее разнообразие начальных сочетаний?
6. Сделайте вывод: чей стиль (Крюкова или Шолохова) ближе к стилю автора «Тихого Дона»?

5. Длина слова в буквах.



Изучите график и ответьте на вопросы:

1. Какая кривая (Крюкова или Шолохова) больше отклонилась от кривой автора «Тихого Дона»?
2. Где наблюдаются наиболее существенные расхождения (для слов какой длины)? С чем это может быть связано?
3. Сделайте вывод: чей стиль (Крюкова или Шолохова) ближе к стилю автора «Тихого Дона»?

6. Лексические спектры/ словарный профиль

Слова, встретившиеся по 1 разу			**20 самых употребительных слов**		
выборка	**% словаря**		**выборка**	**% текста**	
ТД2	40,15		К1	22,88	
ТД4	39,15		К2	21,75	
ТД1	37,78		Ш1	20,64	
Ш2	37,45		ТД2	20,18	
Ш1	36,49		Ш2	19,98	
К2	35,52		ТД1	19,91	
К1	32,02		ТД4	18,70	
Сред. отличие	Ш - ТД	5,28%	**Сред. отличие**	Ш - ТД	3,67%
	К - ТД	13,48%		К - ТД	13,88%

Вопросы.

1. Какие зоны частотного словаря сравниваются в данном случае?
2. Какой писатель (Крюков или Шолохов) по данному параметру ближе к автору «Тихого Дона»?

Методические указания по выполнению заданий для самостоятельной работы.

Самостоятельная работа №1. Статистические инструменты в исследовании объектов филологии.

Цель: сравнить функциональные стили русского языка с помощью количественных методов.

Программное обеспечение: текстовый редактор (Microsoft Office Word или OpenOffice), редактор электронных таблиц (Microsoft Office Excel или OpenOffice Calc).

Материал: тексты из сети Интернет, таблица хи-квадрат, таблица Стьюдента.

1. Выберете в Интернете 2 научные статьи из одной отрасли науки и 2 художественных прозаических произведения разных авторов. Из статей выберете по 3 идущие не подряд фрагмента по 100 слов (для выбора фрагментов нужной длины воспользуйтесь функцией «Сервис» → «Количество слов» в приложении OpenOffice, «Рецензирование» → «Статистика» в приложении Word 2007 или «Сервис» → «Статистика» в приложении Word 2003); из произведений каждого автора выберете по 3 идущих не подряд фрагмента по 100 слов, исключая речь персонажей. Какими способами формирования выборки Вы пользовались? К каким ступеням на шкале однородности по классификации Д.М.Сегала принадлежат выбранные Вами фрагменты?
2. В каждом фрагменте посчитайте количество существительных и заполните в Excel следующую таблицу:

Научный стиль				Художественный стиль			
Статья 1		Статья 2		Автор 1		Автор 2	
№ фрагмента	Число существительных	№ фрагмента	Число существительных	№ фрагмента	Число существительных	№ фрагмента	Число существительных
1		4		1		4	
2		5		2		5	
3		6		3		6	

3. Сравните выборки из научного и художественного стилей по критерию Стьюдента – см. справочник формул (совокупность 1 – тексты научного стиля, совокупность 2 – тексты художественного стиля). Сделайте вывод: является ли количество

существительных стилистически значимым параметром.

4. Проверьте однородность выборок по научному и художественному стилям с помощью критерия согласия и коэффициента вариации[1] (см. справочник формул). Сделайте вывод: выборки однородны/неоднородны. Если выборки неоднородны, предположите, какие могут быть причины этого.
5. Сравните степень однородности по коэффициенту вариации фрагментов внутри каждой из статей и произведений каждого автора[2] со степенью однородности общих выборок по научному и художественному стилям[3] и сделайте вывод: фрагменты внутри 1 статьи (1 автора) более однородны / менее однородны, чем фрагменты из научного (художественного) стиля вообще. Подтвердилась ли на Вашем материале теория Д.М.Сегала о шкале однородности?

Работы сдаются в виде 2 файлов: один (материал) — в формате Word, второй (сама работа) — в формате Excel; оба файла должны быть в папке/архиве с названием: с1_<фамилия студента>_<№ группы>, например, *с1_иванова_211*. Файлы с самой работой должны быть сданы в формате Excel 97/2003; в случае если Вы работаете с другими версиями Microsoft Office или с приложением OpenOffice, воспользуйтесь функцией «Сохранить как» и выберете соответствующий формат в строке «Тип файла». Имя файла задается следующим образом: с1_<фамилия студента>_работа, например, *с1_иванова_работа*. Все данные, расчеты и выводы должны содержаться в сдаваемом файле, все итоговые значения должны быть подписаны. Работы без выводов и расчетов НЕ будут зачтены. В файле с материалом должны содержаться проанализированные Вами фрагменты в виде:

Статья 1.

фрагмент 1.

<текст фрагмента 1>

фрагмент 2.

<текст фрагмента 2>

…...

Статья 2.

…....

[1] т.е. должно получиться 4 числа – критерий согласия по научному стилю, критерий согласия по художественному стилю, коэффициент вариации по научному стилю, коэффициент вариации по художественному стилю

[2] т.е. посчитайте коэффициент вариации по статье 1, по статье 2, по автору 1, по автору 2

[3] общий коэффициент вариации должен был быть посчитан при выполнении задания 4

Файл сдается в формате Word97/2003, название: с1_<фамилия студента>_материал, например, *с1_иванова_материал*.

Работы присылать на электронную почту: gaerven@yahoo.com, в теме сообщения указать «самостоятельная по «Колич. методам»».

Самостоятельная работа №2. Применение методов корреляционного анализа в исследовании объектов филологии.

Цель: с помощью методов корреляционного анализа проверить однородность выборок; установить наличие/отсутствие зависимости между классами лингвистических единиц.

Программное обеспечение: текстовый редактор (Microsoft Office Word или OpenOffice), редактор электронных таблиц (Microsoft Office Excel или OpenOffice Calc), приложение Fonvalue.

Материал: мемуары С. Аллилуевой «Двадцать писем к другу»

1. Из любой главы произведения выберите 7 идущих подряд фрагментов текста по 1000 знаков, включая знаки препинания, но без учета пробелов; Ваш порядковый номер в списке группы соответствует номеру предложения, с которого Вам следует начинать выборку. Сохраните каждый из фрагментов в отдельном файле в формате «Обычный текст» с кодировкой MS-DOS (для OpenOffice: формат “Кодированный текст» → «да» → кодировка Кириллица (DOS/OS2-866/Русский) и разместите файлы в подкаталоге приложения FONVAL «Примеры». Для удобства обработки дайте файлам имена с числовыми индексами. Например, all1.txt, all2.txt, all3.txt и т.д.

Для выделения фрагментов нужной длины воспользуйтесь опцией «Статистика» редактора Word («Сервис» → «Статистика») или OpenOffice (Сервис → Количество слов).

2. Активизируйте приложение FONVALUE. Поочередно открывайте в приложении FONVALUE созданные и сохраненные Вами файлы с фрагментами текста С. Аллилуевой. Используя закладку «Статистика», получите данные об абсолютной частоте «звукобукв» а, о, и, , д, д', м, м', з, ж, с, ш (по А.П. Журавлёву) в каждом из фрагментов в отдельности.

Полученные данные организуйте в виде таблицы, в которой «Ч.» - абсолютная частота, «Р.» - ранг:

Звуко-буквы	Фрагм. 1		Фрагм. 2		Фрагм. 3		Фрагм. 4		Фрагм. 5		Фрагм. 6		Фрагм. 7	
	ч.	р.	ч.	р.	ч.	р.	ч.	р.	ч.	р.	ч.	р.	ч.	р.
«а»														

«о»														
«и»														
«у»														
«д»														
«д'»														
«м»														
«м'»														
«з»														
«ж»														
«с»														
«ш»														

На основе данных таблицы вычислите коэффициенты ранговой корреляции между рангами звукобукв фрагментов № 1 и № 2, фрагментов № 1 и № 3, фрагментов № 1 и № 4 и так далее – см. справочник формул.

Какие выводы относительно однородности фрагментов в парах можно сделать на основе рассмотрения коэффициентов ранговой корреляции?

3. Вычислите среднюю величину абсолютной частоты для каждой из выбранных для анализа звукобукв, подготовьте таблицы для расчета «коэффициента согласия» χ^2 отдельно для каждой из «звукобукв», заполните их и вычислите величину χ^2 - см. справочник формул.

Выборка	$\bar{x}$ (средн.)	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1				
2				
3				
4				
5				
6				
7				
Сумма квадратов разностей $(x_i - \bar{x})^2$				

Какие выводы об однородности выборок из текста С. Аллилуевой позволяет сделать использование «критерия согласия» χ^2?

4. Проверьте однородность выборок по коэффициенту вариации. Сделайте вывод: расхождение частот случайно/не случайно, совокупность однородна/неоднородна

5. Рассчитайте коэффициент корреляции Пирсона (см. справочник формул) для следующих данных

сущ у Грецена	мест у Герцена
150	109
158	106
164	114
168	98
169	109
170	102

171	100
181	68
183	105
183	94
186	85
187	98
190	96
199	85
201	82
201	79
202	76
205	85
213	71
218	73

Сделайте вывод о соотношении существительных и местоимений у А.И. Герцена.

Самостоятельная работа №3. Фонетическое значение слова.

Цель: вычислить фонетическое значение слова.

Материал: данные о фонетическом значении русских звукобукв и их стандартной частотности в русских текстах из эксперимента А.П. Журавлева.

Вычислите фонетическое значение своей фамилии по методике, предложенной А.П. Журавлевым – см. справочник формул. Вы можете взять любую из 25 шкал.

Самостоятельная работа № 4. Ассоциативная сила слова.

Цель: вычислить ассоциативную силу предложенных слов.

Материал: данные Русского ассоциативного словаря под. ред. Ю.Н. Караулова.

Вычислите ассоциативную силу слов в одном из предложенных списков по методике И.Г. Овчинниковой и А.С. Штерн (см. справочник формул). Сделайте вывод о степени стереотипности реакций на данные слова:

a. университет, студент, экзамен
b. мама, дом, семья
c. большой, красивый, маленький
d. деньги, работа, отдых
e. рука, лицо, глаза

Самостоятельная работа №5. Частотные словари.

Цель: выявить особенности картины мира по данным частотных словарей.
Материал: данные Нового частотного словаря русской лексики под ред. О. Н. Ляшевской и С. А. Шарова.

Выпишите первые 30 самых частотных глаголов из частотных словарей художественной литературы, другой нехудожественной литературы и живой устной речи. Распределите глаголы в каждой выборке по тематическим группам. Сопоставьте полученные группы в трех стилях. Сделайте вывод о специфике картины мира, отраженной в текстах разных стилей.

Глоссарий

Абсолютная частота – число появлений изучаемого явления в наблюдаемом отрезке действительности.

Аннотированный (размеченный, структурированный корпус) – корпус с разметкой (см. неаннотированный корпус).

Ассоциативная сила слова (АСС) – способность слова вызывать наиболее предсказуем реакции.

Атрибуция – определение авторства текста; реже – его датировка и т.п.

Богатство рифменного созвучия – количество элементов, составляющих созвучие, т.е. число пар идентичных и тождественных фонем.

Висячий узел – узел, из которого не выходит ни одна стрелка.

Выборка – избранное подмножество объектов генеральной совокупности, обладающее свойством репрезентативности.

Генеральная совокупность – совокупность однородных лингвистических объектов, обладающих изучаемым признаком (т.е. все множество объектов, относительно которых формулируется гипотеза).

Голова частотного словаря – структурная зона ЧС, состоящая из наиболее частотных языковых единиц.

Густота дерева зависимости – число висячих узлов, т.е. таких, из которых не выходит ни одна стрелка; густота – показатель степени распространенности предложения.

Дерево зависимости – способ представления синтаксической структуры предложения в виде графа, узлы которого соответствуют словам предложения, а стрелки – связям между ними.

Динамический корпус – корпус, обновляемый/дополняемый через определенный промежуток времени; позволяет выделить рабочий подкорпус (см. статический корпус).

Длина дерева зависимости – 1. длина дерева в узле X – кратчайшее расстояние по стрелкам от вершины дерева до узла X; длина дерева в целом – максимальное значение длины для данного предложения.

Единица хранения корпуса данных – некоторая совокупность естественно-языковых выражений из проблемной области, которой сопоставляется одно описание на некотором метаязыке, определяемом процедурой формирования корпуса.

Иллюстративный корпус – создается после проведения исследования, чтобы подтвердить и обосновать полученные результаты; включает не все особенности проблемной области, а только факты, релевантные для данного исследования (см. исследовательский корпус).

Интервальная шкала – устанавливается единица измерения; предмету присваивается число, равное количеству единиц измерения в соответствии с количеством измеряемого признака; 0 – условная точка шкалы.

Исследовательский корпус – корпус, предназначенный для изучения различных аспектов функционирования языковой системы; ориентированный на решение широкого круга задач; формируется до проведения конкретного исследования (см. иллюстративный корпус).

Корпус данных – сформированная по определенным правилам выборка данных из проблемной области.

Корпус параллельных текстов – корпус, представляющий собой подмножество текстов на языке-источнике с одним или несколькими подмножествами текстов перевода.

Корпус текстов – вид корпуса данных, единицами хранения которого являются тексты.

Местоположение рифменного созвучия в строке – расстояние от правого края строки до области созвучия.

Неаннотированный (неразмеченный, неструктурированный) корпус – корпус без разметки

Номинальная шкала – распределение объектов по классам

Ординальная шкала – см. порядковая шкала.

Относительная частота – доля изучаемого явления в данной выборке; вычисляется путем деления абсолютной частоты на объем выборки.

Полнота корпуса – требование к корпусу, заключающееся в том, чтобы учитывались все релевантные явления, даже если это не соответствует требованиям пропорционального сужения.

Порог отображения – соотношение между корпусом данных и проблемной областью при пропорциональном сужении.

Порядковая (ранговая, ординальная) шкала – расположение объектов по степени интенсивности признака

Проблемная область – область реализаций языковой системы, содержащая феномены, подлежащие лингвистическому описанию.

Проективность – предложение считается проективным, если при представлении его структуры в виде дерева зависимости стрелки не пересекают друг друга.

Разметка – приписывание текстовым единицам в корпусе набора параметров, определяемых целью создания корпуса (напр., морфологическая, синтаксическая, семантическая, акцентологическая и т.п. разметка).

Ранговая шкала – см. порядковая шкала.

Репрезентативность – способность выборки отражать все свойства генеральной совокупности, релевантные для данного типа лингвистического исследования, в определенной пропорции, определяемой частотой явления в генеральной совокупности.

Статический корпус – корпус, не пополняемый регулярно (напр., авторский корпус)

Степень гнездования дерева зависимости – максимальное число стрелок, накрывающих узел и не имеющих общих краев; показатель того, насколько далеко находятся синтаксические связи

Степень дистантизации дерева зависимости – максимальное число узлов между двумя синтаксически связанными узлами.

Стилеметрия – измерение стилистических явлений с целью упорядочивания и систематизации текстов и их частей.

Тело частотного словаря – структурная зона ЧС, включающая знаменательные слова с абсолютной частотой не равной единице.

Фонетическое значение – соответствие символико-содержательного аспекта звуковой формы слова его понятийному значению.

Функциональные стили языка – типы его функционирования, соответствующие различиям социальной практики коллектива и отличающиеся существенными различиями вероятностей языковых единиц и их категорий, достаточными для их совокупного качественного опознавания людьми на интуитивном уровне восприятия речи.

Хвост частотного словаря – структурная зона ЧС, состоящая из наиболее редких слов, абсолютная частота которых равна 1.

Частотный словарь – словарь особого типа, представляющий собой список языковых единиц с указанием при каждой частоты ее употребления в совокупности текстов определенной длины.

Ширина дерева зависимости – 1. ширина дерева в узле X – число стрелок, выходящих из этого узла; 2. ширина дерева в целом – максимальное значение ширины для данного предложения; ширина дерева отражает полноту представления ситуации.

Шкала отношений – аналогична интервальной шкале, однако здесь 0 – не условная точка шкалы, а полное отсутствие признака.

Экономичность корпуса – требование к корпусу, заключающееся в том, что корпус должен быть существенно меньше проблемной области.

Справочник формул.

Ассоциативная сила слова: $v = \frac{1}{H}$; $H = -\sum p_i * \log_2 p_i$, где p_i – относительная частота реакции R_i

Закон Ципфа: $p_r = \frac{k}{r^\gamma} = k * r^{-\gamma}$, где r – ранг языковой единицы при упорядочивании по убыванию частотности, p_r – относительная частота языковой единицы с рангом r, k, γ – константы ($k \approx 0{,}1$; $\gamma \approx 1$)

Коэффициент (ранговой) корреляции Спирмена: $\rho = 1 - \frac{6 * \sum d^2}{n^3 - n}$, где d – разность соответствующих рангов, n – максимальный ранг в двух совокупностях

Коэффициент вариации: $v = \frac{\sigma}{\bar{x}} * 100$, где σ – среднеквадратичное отклонение, $\bar{x}$ – средняя выборочная частота.

Коэффициент корреляции Пирсона: $r = \frac{\sum a_i b_j}{\sqrt{\sum a_i^2 * \sum b_j^2}}$, где a_i, b_j – отклонения от средних выборочных частот

Критерий согласия – см. хи-квадрат критерий.

Критерий Стьюдента: $t = \frac{\overline{x_1} - \overline{x_2}}{s_{1,2}} * \sqrt{\frac{k_1 * k_2}{k_1 + k_2}}$, где $\overline{x_1}$, $\overline{x_2}$ – средние выборочные частоты в 1й и 2й совокупностях, k_1, k_2 – число выборок в 1й и 2й совокупностях, $s_{1,2}$ – среднеквадратичное отклонение по двум совокупностям.

Среднеквадратичное отклонение (по 2м совокупностям):

$s_{1,2} = \sqrt{\frac{\sum (x_{i1} - \overline{x_1})^2 + \sum (x_{i2} - \overline{x_2})^2}{k_1 + k_2 - 2}}$, где x_{i1}, x_{i2} – конкретные выборочные частоты в 1й и 2й совокупностях, $\overline{x_1}$, $\overline{x_2}$ – средние выборочные частоты в 1й и 2й совокупностях, k_1, k_2 – число выборок в 1й и 2й совокупностях

Среднеквадратичное отклонение: $\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{k - 1}}$, где x_i – конкретная выборочная частота, k – число выборок.

Средняя выборочная частота: $\bar{x} = \frac{\sum x_i}{k}$, где x_i – конкретная выборочная частота, k – число выборок.

Степень близости ассоциативных полей: $w = \sum_{i=1}^{K} \frac{\frac{f_{i1}}{N_1} + \frac{f_{i2}}{N_2} - \left| \frac{f_{i1}}{N_1} - \frac{f_{i2}}{N_2} \right|}{2}$, где K – общее число различных ассоциаций в обеих группах; f_{i1}, f_{i2} – частоты i-той ассоциации в первой и второй группах, N_1 и N_2 – численности 1й и 2й выборок, т.е. число испытуемых

Степень близости говоров (по Н.Н. Пшеничновой): $t_{ij} = \frac{1}{n} \sum_{r=1}^{n} w_r$, где n – общее число признаков, r – номер признака, w_r – показатель близости говоров с номерами i и j по признаку r; если в обоих говорах есть признак r, $w_r = \frac{S - k_r}{k_r}$; если в обоих говорах нет признака r, $w_r = \frac{k_r}{S - k_r}$, где S – общее число говоров совокупности, k_r – число говоров с признаком r в данной совокупности; если в 1 говоре признак есть, а в другом нет, $w_r = -1$

Степень близости языков (по А.Я. Шайкевичу): $S_{AB} = \frac{100 * \sum x_i(AB)}{N_A}$, где x_i – общий признак в языках A и B, N_A – общее число выраженных в языке А признаков.

Фонетическое значение звукобуквы: $x_i = \frac{\sum r * f_r}{\sum f}$, где r – ранг, f_r – число испытуемых, приписавших данный ранг данной звукобукве по данной шкале.

Фонетическое значение слова: $s_t = \frac{\sum x_i k_i}{\sum k_i}$, где x_i – фонетическое значение i-той звукобуквы по данной шкале, $k_i = \frac{P_{\max}}{P_i}$, где P_i – стандартная частота i-той звукобуквы в русских текстах, P_{max} – максимальная частота звукобуквы в данном слове (для 1й буквы k_i умножается на 4, а для ударной гласной – на 2)

Фонетическое значение текста: $z = \sum y'_i * x'_i$, где y'_i – степень отклонения частоты i-той звукобуквы в данном тексте от ее стандартной частоты в русских текстах (учитываются только те звукобуквы, для которых $y_i > 1$), x'_i – фонетическое значение i-той звукобуквы (учитываются только те звукобуквы, для которых $y_i > 1$); $y_i = \frac{P_i(k) - P_i(N)}{\sigma_{P_i(N)}}$, где $P_i(N)$ – нормальная (стандартная) частота i-той звукобуквы, $P_i(k)$ – частота i-той звукобуквы в данном произведении, $\sigma_{P(N)}$ – среднеквадратичное отклонение.

Хи-квадрат критерий (критерий согласия): $\chi^2 = \frac{\sum (x_i - \bar{x})^2}{\bar{x}}$, где x_i – конкретная выборочная частота, $\bar{x}$ – средняя выборочная частота.

Литература.

Алексеев П.М. Частотные словари. СПб., 2001.

Арапов М.В. Квантитативная лингвистика. М., 1988.

Баранов А.Н. Введение в прикладную лингвистику. М., 2001.

Вероятностное прогнозирование в речи. М., 1971.

Гаспаров М.Л. Точные методы анализа грамматики в стихе // Славянское языкознание. XII Международный съезд славистов. Краков. 1998 г. Доклады российской делегации. М., 1998.

Головин Б.Н. Язык и статистика. М., 1971.

Журавлев А.П. Фонетическое значение. Л., 1974.

Зиндер Л.Р. О лингвистической вероятности // ВЯ, 1958. № 2.

Квантитативная лингвистика и семантика. Новосибирск, 1999.

Мартыненко Г.Я. Основы стилеметрии. Л., 1988.

Налимов В.В. Вероятностная модель языка: о соотношении естественных и искусственных языков. М., 1979.

Наследов А.Д. Математические методы психологического исследования. СПб., 2004.

Носенко И.А. Начала статистики для лингвистов. М., 1981.

Овчинникова И.Г., Штерн А.С. Ассоциативная сила русского слова // Психолингвистические проблемы фонетики и лексики. Калинин, 1989.

Пиотровский Р.Г, Бектаев К.Б., Пиотровская А.А. Математическая лингвистика. М, 1977.

Пиотровский Р.Г. Информационные измерения языка. Л., 1968.

Поликарпов, А.А. Проблемы квантитативно-системной лингвистики // Учен. зап. Тарт. ун-та. Тарту, 1989. - Вып. 872.

Поликарпов, А.А.; Тулдава, Ю.А. Частотные словари и опыт их использования // Учен. зап. Тарт. гос. ун-та. 1987. - Вып. 774.

Пшеничнова Н.Н. Типология русских говоров . М., 1996.

Русский язык по данным массового обследования: Опыт социально-лингвистического изучения. М., 1974.

Сегал Д.М. Основы фонологической статистики. М., 1972.

Статистика речи и автоматический анализ текста / Отв. ред. Р.Г. Пиотровский. Л., 1980.

Угланова И.А. Субъективная оценка частоты слова и ее категоризация. Пермь, 2005.

Фрумкина Р.М. Статистические методы изучения лексики. М., 1964.

Хьетсо Г., Густавссон С., Бекман Б., Гил С. Кто написал «Тихий Дон»? (Проблема авторства «Тихого Дона»). М., 1989.

Шайкевич А.Я. Гипотезы о естественных классах и возможность количественной таксономии в лингвистике // Гипотеза в современной лингвистике. М., 1980.

Шайкевич, А.Я. О Статическом словаре языка Достоевского // Рус. яз. в науч.

Шепелева С.Н., Петров В.М. К проблеме описания эволюционных параметров русской рифмы // Проблемы структурной лингвистики – 1978. М., 1981.

Языковая норма и статистика. М., 1977.

Якубайтис Т.А., Скляревич А.Н. Вероятностная атрибуция типа текста по нескольким морфологическим признакам. Рига, 1982.