ЛАБОРАТОРНАЯ РАБОТА 2

«Статистические параметры языка. Количественные оценки»

Статистическая лингвистика – это дисциплина, изучающая количественные закономерности естественного языка, проявляющиеся в текстах. В основе лежит предположение, что некоторые численные характеристики и функциональные зависимости между ними, полученные для ограниченной совокупности текстов, характеризуют язык в целом или его функциональные стили (публицистический, научный, художественный и т.п.). Практически важной и наиболее изученной числовой характеристикой является относительная частота употребления различных лингвистических единиц (букв, фонем, слогов, слов, синтаксических конструкций), их классов (например, гласных, согласных, частей речи) и сочетаний (например, последовательностей из n букв). Данные о частоте слов (иногда словосочетаний) отражаются в частотных словарях. Статистическая лингвистика изучает также зависимости между частотой и длиной слова (в числе слогов), числом его значений и возрастом. Накопленные данные используются для выявления особенностей стиля отдельных авторов, атрибуции текстов, дешифровки исторических письменностей, для решения задач стенографии, теории связи, а также информатики. Статистическая лингвистика при получении численных характеристик использует методы математической статистики и некоторые методы теории информации (для определения энтропии и избыточности языка, а для установления связи между наблюдаемыми характеристиками и выбора наиболее существенных из них — метод математических моделей, базирующихся на понятиях теории вероятностей и математической лингвистики (рис. 1).

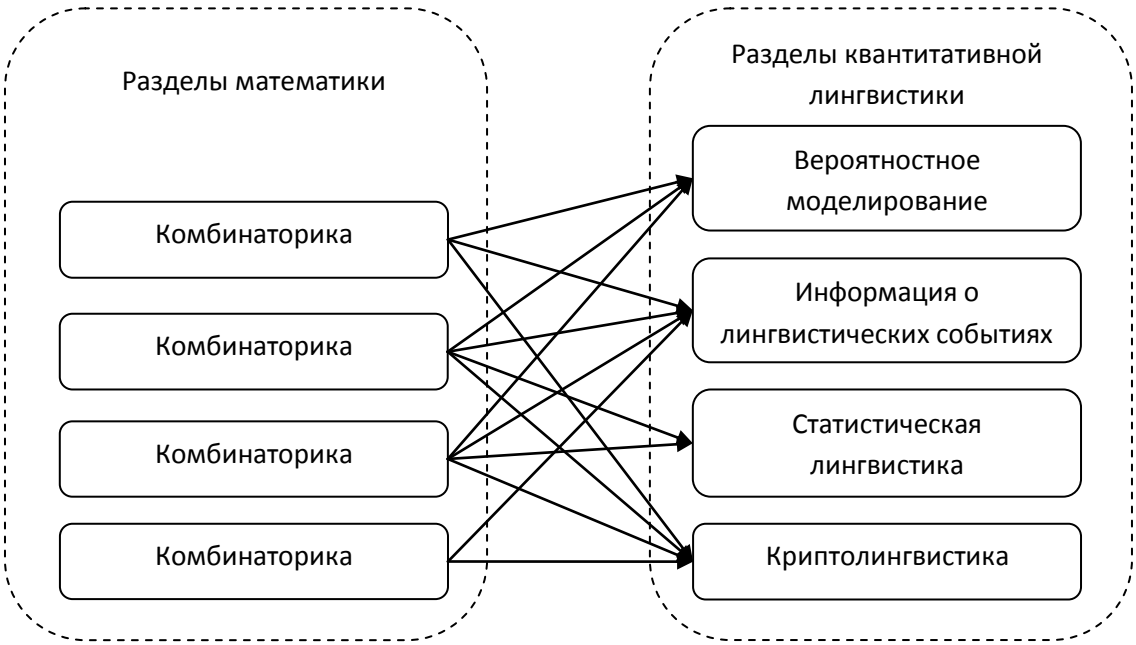

Рис. 1 Связи между методами разделов математики и разделом статистической лингвистики

Лингвометрия – это область прикладной лингвистики, которая определяет, измеряет и анализирует количественные единицы разных уровней языка. Используя аппарат математической лингвистики, лингвометрия помогает в решении таких задач как создание и сравнение словарей, автоматическое создание словарей, тезаурусов, создание систем стенографии, автоматическое определение языка, информационный поиск и т.д.

Статистические параметры языка

Каждый язык имеет собственные статистические параметры, и знание частоты появления букв и их словосочетаний (биграмм, триграмм, четырехграмм) определенного языка дает возможность автоматически ее идентифицировать. Например, для украинских текстов было определено, что статистическими параметрами стилей можно считать частоты гласных и согласных, пропуски между словами, а также мягкие и сонорные группы согласных. Для украинского языка данные, которые содержат частоты гласных и согласных, частоты групп сонорных, звонкие и глухие согласные и их оценки, а также частоты использования букв языка приведены в таблице 1.

Покажем, как выполнить оценку (украинского или российского языка) отрывка текста с помощью определенного эталона – например, данных про частотность букв украинского языка.

Рассмотрим два отрывка текста на украинском и российском языках, представленные в формате, где буквы расположены по уменьшению частот их появления в отрывке, а разделение на маленькую и большую букву не делается. Найдем тип корреляции частот букв отрывков и эталона, результаты, которые подтверждены выводами, представим в графическом виде.

В таблице 1 для удобства внесем такие данные: частотность использования букв украинского языка, абсолютные и относительные частоты использования букв в исследованиях. Отрывок 1 содержит 556 символов, отрывок 2 – 541 символ. Понятие «інші» в столбце букв содержат аутентичные буквы для украинского (ї, є, і) и российского языков (ё, ы, ъ, э), что дает возможность достичь определенной независимости во время анализа. Полученные результаты графически представим на рис. 2

Таблица 1.

Частотность появления букв в эталоне и в исследуемых отрывках

Літера	Частотність вживання літер української мови (еталон)	Абсолютна частота літер в Уривку 1	Відносна частота вживання літер в Уривку 1	Абсолютна частота літер в Уривку 2	Відносна частота вживання літер в Уривку 2
« »	0,133	80	0,14	82	0,15
о	0,082	37	0,07	41	0,08
а	0,074	43	0,08	31	0,06
н	0,068	33	0,06	30	0,06
и	0,054	27	0,05	27	0,05
в	0,047	29	0,05	19	0,04
т	0,046	25	0,04	20	0,04
е	0,038	26	0,05	45	0,08
р	0,036	15	0,03	16	0,03
с	0,033	22	0,04	27	0,05
м	0,031	10	0,02	13	0,02
к	0,031	22	0,04	20	0,04
л	0,028	17	0,03	30	0,06
д	0,028	16	0,03	4	0,01
у	0,025	19	0,03	14	0,03
п	0,025	11	0,02	21	0,04
я	0,024	15	0,03	6	0,01
з	0,018	9	0,02	8	0,01
б	0,016	7	0,01	5	0,01
ч	0,015	5	0,01	11	0,02
г	0,012	4	0,01	6	0,01
ю	0,012	2	0,00	2	0,00
б	0,011	7	0,01	5	0,01
х	0,01	4	0,01	7	0,01
ц	0,009	7	0,01	1	0,00
ж	0,007	3	0,01	7	0,01
й	0,007	4	0,01	6	0,01
ш	0,005	3	0,01	2	0,00
щ	0,004	3	0,01	1	0,00
ф	0,003	1	0,00	0	0,00
інші	0,0605	51	0,09	34	0,06

На рисунке 2 показано, что частота появления букв в отрывках не дает убедительного ответа на вопрос, кокой из отрывков написан на украинском языке. Однако, можно видеть резкий скачок относительной частоты появления буквы «е» для отрывка 2 относительно эталонных значений, потому будем считать, что с большей вероятностью на украинском языке написан отрывок 1.

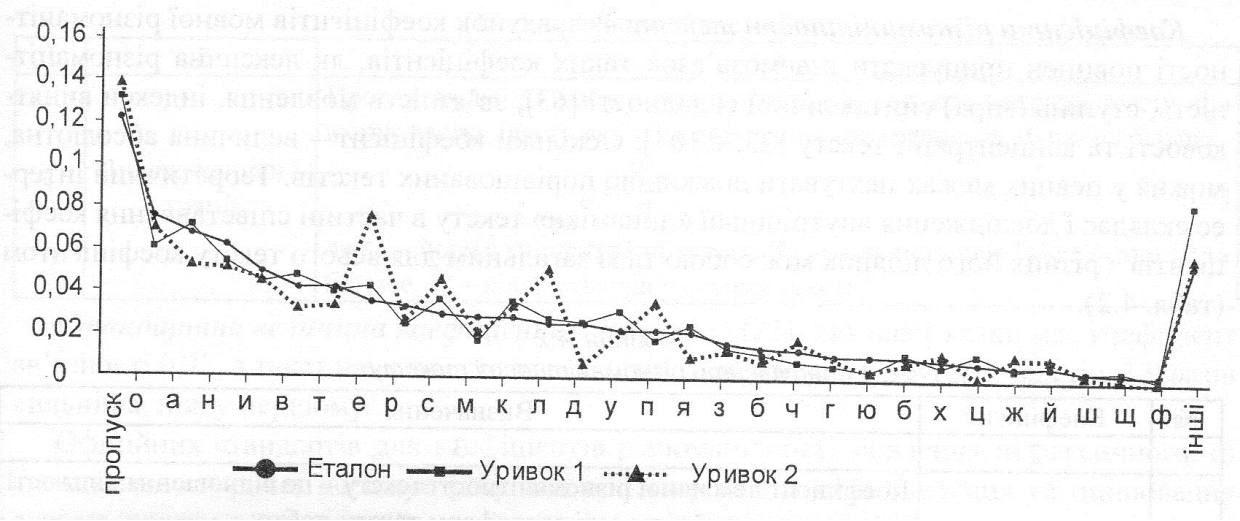

Рис. 2 Графическое отображение относительных частот появления букв в эталоне и в исследуемых отрывках

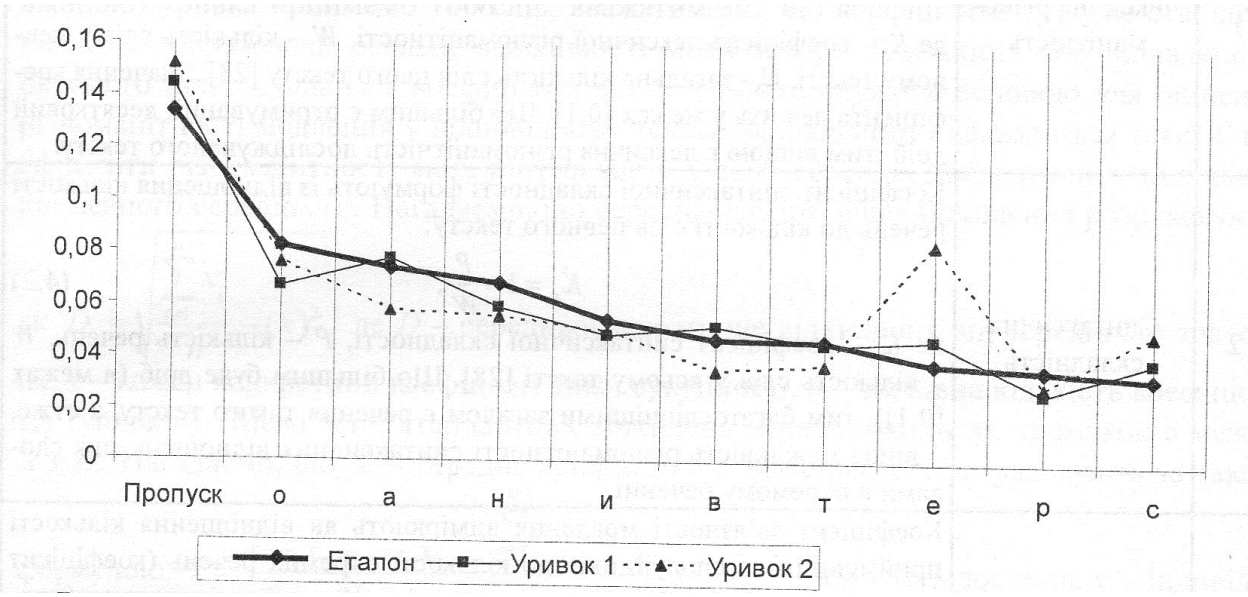

Рис. 3 Графическое отображение относительных частот появления десяти наиболее частотных символов в эталоне и в исследуемых отрывках

Количественные оценки речи

Коэффициенты разнообразия текста. Расчет коэффициентов языкового разнообразия должен допускать взаимосвязь таких коэффициентов, как лексическое разнообразие, степень синтаксической сложности, связность речи, индекс исключительности и концентрации текста. Так как коэффициент – это величина абсолютная, то можно в определенных границах пренебрегать длинной сравниваемых текстов.

Таблица 2. Коэффициенты разнообразия текста

№	*Коэффициент*	*Определение*
1	Лексическое разнообразие	Коэффициент лексического разнообразия текста – это отношение количества слов к общему количеству словоформ текста, т.е. $$K = \frac{W}{N},$$ где K – это коэффициент лексического разнообразия, W – количество слов в определенном тексте, N – общее количество слов этого текста. Значение коэффициента лежит в пределах от 0 до 1.
2	Синтаксическая сложность	Коэффициент синтаксической сложности формируют как отношение количества предложений к количеству слов определенного текста: $$Ks = 1 - \frac{P}{W},$$ где Ks – это коэффициент синтаксической сложности, P – количество предложений, W – количество слов во всем тексте. Значение коэффициента лежит в пределах от 0 до 1.
3	Коэффициент связности речи	Коэффициент связности речи измеряется как отношение количества предлогов и союзов к количеству отдельных предложений (коэффициент равен 1, если в одном предложении есть три связных элемента): $$Kz = \frac{Z + S}{3P},$$ где Z – это количество предлогов, S – количество союзов, P – отдельных предложений.
4	Индекс исключительности	Индекс исключительности характеризует вариативность лексики, т.е. часть текста, которую занимают слова, что встретились 1 раз и вычисляется следующим образом: $$I = \frac{W_1}{W},$$ где W_1 – это количество слов с частотой 1, W – количество слов во всем тексте.
5	Индекс концентрации	Противоположным к индексу исключительности есть индекс концентрации текста, что показывает часть текста, которую занимают слова, которые появились 10 раз и больше: $$Ik = \frac{W_{10}}{W},$$ где W_{10} – это количество слов с частотой 10 и больше, W – количество слов во всем тексте.

В результате множества экспериментов было определено, что текст сказки имеет коэффициент связности 0,77, а текст научной статьи – 3,0.

Официальных стандартов для коэффициентов разнообразия речи ни лексического, ни синтаксического уровня не существует. Ориентиром для сопоставления и оценивания какого-нибудь текста в однородной группе текстов может служить среднестатистическая

норма величины коэффициента для равных по длине отрывков. Оптимальным размером отрывка принято считать 100 слов, считается, что коэффициенты тут же стабилизируются, отображая реальные особенности языка автора.

Задание на лабораторную работу

1) Используя формулы, представленные выше, оцените произвольные отрывки художественного и научного текста с помощью коэффициентов разнообразия текстов.
2) Используя статистические параметры языка, сравните два текста (на английском и немецком языках). В качестве эталона возьмите текст – «Алиса в стане чудес» Льюиса Кэрролла.