УДК 1(03):801

ББК А5:87

Б 26

Philosophia magistra vitae

ЛИНГВИСТИКА И ФИЛОСОФИЯ. Часть 3.

КАТЕГОРИЯ КОЛИЧЕСТВА

Бартков Б.И.

Доцент

Дальневосточное отделение Российской академии наук

Владивосток, Россия

Аннотация. Приводятся результаты критического анализа основных работ, посвящённых применению математических методов для точного описания звукозначащих элементов речи (звуков, морфов, слов, непредикативных и предикативных словосочетаний) и языка (фонем, морфем, лексем, фразем, паремием) и их взаимосвязи в пределах подсистем (фонемной, морфемной, лексемной, синтаксемной (то есть фраземной, паремиемнолй) и между ними во времени и пространстве. При описании математических методов использовались иллюстрации из работ ряда лингвистов, в том числа и автора. Показано, что использование категории количество в дополнению к категории качества даёт положительные результаты.

Ключевые слова: лингвистика, философия, категории качества и количества, термины система, речь, язык и их элементы.

LINGUISTICS AND PHILOSOPHY: Part 3.

CATEGORY OF QUANTITY

Bartkov B.I.

Associate Professor

Far-Eastern Branch of the Russian Academy of Sciences

Vladivostok, Russia

Abstract. Results of critical analysis of major works devoted to application of quantity category to precise description of sounding and meaning elements of speech (sound, morph, word, predicative and non-predicative word-combinations) and language (phoneme, morpheme, lexeme, phraseme, paremieme) in time and space. While describing mathematical methods, illustrative examples have been taken from a few authors including the author.

Keywords: linguistics, philosophy, categories of quantity and quality, system, speech, language and their elements.

Начиная изучать философию, лингвистику или математику, необходимо ознакомиться с научно корректными дефинициями этих наук.

Философия – *это наука о наиболее общих закономерностях изменения в пространстве и времени таких материальных сущностей (систем), как бытие, человеческое общество, мышление (сознание) и язык [а также музыка и ИЗО], состоящих из двуплановых гетерогенных элементов. объединяющихся в подсистемы, характеризующиеся структурой и* функцией [11].

Философский энциклопедический словарь даёт следующие дефиниции категорий (авторы А.Г. Спиркин, М.Г. Ярошевский) [42, 251].

«КАТЕГОРИИ (от греч. Kategoria – высказывание, обвинение; признак) в *философии*, предельно общие, фундаментальные понятия, отражающие наиболее существенные, закономерные связи и отношения реальной действительности и познания. Будучи формами и устойчивыми организующими принципами процесса мышления, категории воспроизводят свойства и отношения бытия и познания во всеобщей и наиболее концентрированной форме» [42, 251].

«Впервые учение о категориях было систематически изложено в трактате Аристотеля «Категории…». Составленная Аристотелем таблица включала такие категории: сущность (субстанция), количество, качество, отношение, место, время, положение, состояние, действие и страдание» [42, 251].

Заметим, что Аристотель ввёл 10 категорий.

«Кант рассматривал категории как априорные рассудка, характеризующие не мир «вещей в себе», а познающего субъекта, структуру его мышления.

У Канта категории делятся на следующие разряды: качество (реальность, отрицание, ограничение), количество (единство, множество, цельность), отношение (субстанция и свойство, причина и действие, взаимодействие), модальность (возможность, невозможность, действительность и недействительность, необходимость и случайность) [43, 251].

В философии «КОЛИЧЕСТВО – это категория, которая отображает общее и единое в вещах и явлениях, характеризуя их с точки зрения относительного безразличия к конкретному содержанию и качественной природе» [42, 263].

Философы считают, что «КАЧЕСТВО – это категория, выражающая неотъемлемую от бытия объекта его существенную определённость, благодаря ко торой он является именно этим, а не иным объектом» [42, 263].

В лингвистике, начиная со знаменитого швейцарского лингвиста Ф. де Соссюра, различают две большие подсистемы: «речь» и «язык».

«Речь» - это то, что мы говорим (и, слава Кириллу и Мефодию, пишем), иначе говоря, речевой поток [42, 263].

«Речь» - это поток взаимосвязанных звуковых и значащих знаков (звуков, морфов, слов, непредикативных и предикативных словосочетаний), передающих информацию (то есть единицы мышления: понятия, суждения и умозаключения).

«Язык – это система обобщённых речевых знаков (фонем, морфем, лексем, фразем и паремием) и правил их употребления в речи, служащая для передачи элементов мышления (понятий, суждений и умозаключений)» [11].

Попросту говоря, «язык» - это грамматика и словарь.

Важно знать, что каждая из этих двух подсистем построена из разных, но генетически соотносимых единиц (знаков) (Табл. 1).

Таблица 1. Единицы речи и языка

Речь	Язык
Звук	Фонема
Морф	Морфема
Слово	Лексема
Словосочетание (непредикативное)	Фразема
Словосочетание (предикативное)	Паремиема

Примечание. Отметим, что наименования всех «языковых» единиц (знаков) имеют в своём составе специальный «лингвистический» суффикс: –ема.

МАТЕМАТИКА (лат. Mathematica < гр. mathematike, от mathema = познание, наука) – 1) наука, изучающая количественные отношения и пространственные формы действительного мира [16, 343].

Философские категории «качества» и «количества» широко использовались при описании различных систем со времён пифагорейцев [42, 726-732].

Древние греки измеряли пирамиды и земельные участки, из чего в рамках натурфилософии возникла «геометрия» Эвклида.

«Математическая пословица» «Пифагоровы штаны во все стороны равны» свидетельствует о популярности математики, в частности, геометрии, среди этого великого народа.

Галилео Галилей (XVI – XVII вв.) ввёл в науку понятие о «законе природы» и показал, что все явления можно описывать **количественно** (чему помогло открытие Иоганном Кеплером законов движения планет вокруг Солнца) [42, 726-732]

Позднее категория *«количества»* стала использоваться в астрономии, механике, технике, физике такими учеными, как *Галилео Галилей, Джордано Бруно, Исаак Ньютон, Рене Декарт и т.д.* [42, 726-732]

Что же касается лингвистики, то известный языковед И. Бодуэн де Куртенэ в 1904 г. писал следующее: «Поскольку в языкознании применяются количественные понятия, желательно тоже знание математики, не только низшей, но и высшей… Если языкознание должно в самом деле стать точной наукой, оно должно научиться орудовать абстракциями, общими понятиями, наподобие того, как это имеет место в настоящих точных науках» [15, 173].

Одними из первых лингвистов, использовавших арифметические подсчеты для описания частотности употребления звуков в немецком [1] и французском [27] (в выборках по 1000 звуков). С современной точки зрения такой объём выборок является малым для того, чтобы получить статистически достоверные результаты, но это были первые количественные оценки частотности звуков – в этом их ценность!

Позднее количественное описание звуков русской устной речи, проделал [28]. Он проанализировал 10 000 звуков и показал, что по частотности (в порядке убывания) они распределяются следующим образом: [А], [Ъ], [Т], [Н], [И], [Эи], [С], [Й], [У], [О], [Л], [Р], [В], [Э], [К], [Д], [П], [М], [Ш], [Ы], [Б], [Ч], [З], [Х], [Ц], [Г], [Ф], [Ж].

Интересно, что он также описал отношение так называемых музыкальных тонов (гласные, сонорные и звонкие шумные) к шумам (глухие шумные согласные) = 74,5% : 25,5%; гласных к согласным = 46,5% : 53,5% и ряд других соотношений т. д. **[19**].

Отметим факт использования «филологического материала» в математических целях академиком А.А. Марковым [21]. Он отсчитал 20 000 букв подряд из первой главы и шестнадцати строк второй главы «Евгения Онегина» для нахождения вероятности появления гласной после гласной или согласной, а также согласной после согласной или гласной. В результате он открыл так называемые «цепи Маркова».

Позднее Н.А. Морозов [22] попробовал использовать математические подсчёты в качестве средства для отличия плагиата от истинных произведений известных писателей. В качестве предмета описания он подсчитал все служебные части речи в первой тысячи ряда произведений таких авторов, как Карамзин, Пушкин, Загоскин, Гоголь, Л. Толстой и Тургенев. Полученные величины частот слов каждой части речи были представлены в виде графиков, названных автором «лингвистическими спектрами».

В результате оказалось, например, что графики наиболее употребительных предлогов у Пушкина и Тургенева оказались одинаковыми, Но делать какие-нибудь статистически достоверные выводы объём выборки не позволяет, хотя эти «стилеметрические этюды», как их называет автор, представляют определённую лингвистическую ценность. Кстати, он первым показал, что предлоги В, НА, С являются самыми частотными.

Эта работа бала первой работой по применению методов математической статистики в языкознании, как считает [19].

В основном, авторы описывают частотности различных фонетических единиц – звуков или букв, или их сочетаний («сгущений» по словам авторов). Например, оказалось, что в славянских, немецком, французском и английском языках наибольшее количество согласных имеет болгарский язык, а наименьшее – французский. Или среди славянских языков русский обладает наибольшим числом гласных звуков. Что касается частотности употребления разных звуков в различных языках, то в русском, немецком и итальянском наибольшую частотность имеет сонорный носовой [n], а звук [r] наиболее частоте описанных в н во французском языке; а звук [s] является наименее частотным во всех описываемых в работе языках. Ценным здесь является то, что полученных характеристики звуков отнюдь не являются очевидными!

Кроме звуков авторы описывают частотности употребления некоторых морфем, предлогов и других частей речи. Так, они подсчитали, что наиболее частотными в русском языке являются следующие предлоги (в порядке убывания частотности): В, НА, С, ЗА, К, ПО, ОТ, ИЗ, ОБ, У.

Следует отметить, что при описании префиксов русского языка было обнаружено, что с увеличением фонетической длины префиксов уменьшается их продуктивность, то есть их способность сочетаться с разными основами.

Авторы также подсчитали, что наиболее частотными корневыми морфемами являются следующие: ВОД, РЯД, ПИС, ДАТ, САД, КЛАД, КАЗ, ВЕД, КИД.

Авторами были получены интересные данные о соотношении частей речи в текстах разных языков. Так. количество существительных, прилагательных и глаголов относится ко всем остальным частям речи как 1 к 3 в русском, литовском, английском, немецком, французском (индоевропейская семья) и древнееврейском (семито-хамитская семья). Эти результаты представляют интерес для лингвистов, занимающихся типологией языков мира.

Это лишний раз подтверждает верность методики изучения не просто каких-либо интуитивно отобранных слов, а наиболее высокочастотных.

Поразительно, но ещё в самом конце XIX века был опубликован
первый частотный словарь немецкого языка, автор которого [55], проанализировав выборку невероятно большого объёма (11 млн. словоупотреблений - СУ) и подсчитал частотность 70 000 слов. Правда, грамматические формы не были сведены к исходным (как это принято в наше время), но всё же этот словарь до сих пор не потерял своего значения.

Главной идеей автора было получение рационального распределения клавишей для появившихся в то время пишущих машинок.

Однако позднее появилось новое поле применения этого словаря – рациональная методика обучения немецкому языку, заключающаяся в том, что сначала изучаются самые частотные слова, затем менее частотные и т.д. Это позволило резко повысить скорость изучения иностранного языка.

Заметим, что в пятидесятые годы XX столетия было подсчитано, что, например, в английском языке 1100 самых частотных слов покрывают 85% любого текста!

Это лишний раз подтверждает верность методики изучения не просто каких-либо интуитивно отобранных слов, а наиболее высокочастотных.

В 1934 году вышел четырёхязычный словарь [48], содержащий список тысячу самых частотных слов английского, французского, немецкого и испанского языков, который можно использовать для рационального и быстрого изучения лексики этих языков.

[54]. Он содержал всего 5 000 слов (объём выборки был равен 1 млн. словоупотреблений).

И, наконец, Частотный словарь русского языка [47], содержащий 40 000 слов с указанием частотности каждого слова (выборка исходного материала равнялась 1 млн. словоупотреблений, что в наше время считается достаточной для получения надежных результатов.)

В 1967 году вышел частотный словарь современного английского языка, содержащий 50 тыс. слов с указанием частотности каждого слова (объем выборки равен 1 млн. словоупотреблений [56].

Заметим, что в пятидесятые годы XX столетия было подсчитано, что, например, в английском языке 1100 самых частотных слов покрывают 85% любого текста!

Это лишний раз подтверждает верность методики изучения не просто каких-либо интуитивно отобранных слов, а наиболее высокочастотных.

Известный английский методист и лексикограф Ю. Торндайк в двадцатых годах XX века выпустил первый частотный словарь английского языка, содержащий 10 тыс. самых частотных слов, затем словарь, содержащий 20 тыс. слов, а в 1944 г. был опубликован частотный словарь, содержавший 30 тыс. наиболее частотных английских слов (объём выборки равнялся в 18 млн. словоупотреблений (СУ) [65].

Однако все эти три словаря имели довольно специфический вид, а именно: частотность отдельных слов не указывалась, а сами слова были сгруппированы по 500 штук: то есть в самом начале этого словаря стояли 500 наиболее высокочастотных слов (расположенных по алфавиту). Затем шли 500 менее частотных слов, также распложенные по алфавиту и так далее.

Этим частотным словарём [65] заинтересовался американский лингвист К. Ципф [68] (заметим, что эту немецкую фамилию американцы по простоте душевной читают и произносят как Зипф). Он обратил внимание на то, что все высокочастотные слова, попавшие в 1-ю полутысячу, в среднем являются фонетически более короткими, чем в следующей 2-й полутысяче, содержащей менее частотные слова, и так далее. То есть он экспериментально обнаружил закономерность: «чем *выше частотность слов, тем они короче фонетически»*.

Будучи лингвистом любознательным, он решил посмотреть в толковом словаре, сколько значений имеется у высокочастотных слов и сколько – у низкочастотных. И он экспериментально обнаружил ещё одну важную лингвистическую закономерность, а именно: «*Чем выше частотность слова, тем оно многозначнее» (*то есть полисемичнее*)*.

В лингвистике давно существуют две важные проблемы:

(1) почему слова, которые всегда впервые возникают с одним значением, затем приобретают второе значение, третье и т. д., становясь многозначными;

(2) почему многие «длинные» слова становятся «короткими» (особенно это важно в отношении английского языка).

Например, английское слово boatswain (боцман) вначале читалось как [‘boutswein], но в наше время произносится (читается) как [‘bousn]; слово cupboard (буфет) произносилось как [‘k/\pbo:d] , а сейчас произносится как [‘k/\b*d] (*звёздочка обозначает нейтральный звук).

Но это открытие осталось незамеченным лингвистами, так как было опубликовано в психологическом журнале [68] (видимо, редакторы лингвистических журналов не нашли ничего интересного в его выводах – подумаешь, короткие и многозначные слова являются высокочастотными! Эка невидаль, вероятно, подумали они…).

Но в 1954 г. французский лингвист Пьер Гиро [50], знакомый с математической статистикой, опубликовал книгу: «Статистические характеристики слов французского языка» [50]. Известно, что ещё в 1930 году в Нью Йорке был опубликован «Частотный словарь французского языка» [66]. Анализируя этот словарь, Пьер Гиро пришел к выводу, что частотность – это очень важная характеристика слова. Оказалось, что высокочастотные слова с общим значением встречаются чаще, чем с узким, что высокочастотные слова относятся к наиболее древним [19]. Он ввёл в лингвистику «коэффициент поэтичности» $K = (V-P) / (V + P)$, где V – частота слова в стихах, P – частота слова в прозе. С помощью этого коэффициента он описывал творчество таких французских поэтов, как Бодлер, Валери, Клодель, Рэмбо.

Отметим, что П. Гиро [50] вывел эмпирическую формулу, связывающую частотность (f) и фонетическую длину слова (L) во французском языке

$$L=C_0 / \lg f,$$

где f – частота употребления соответствующего слова в речи, C_0 – коэффициент пропорциональности.

Интересна также эмпирическая формула, выведенная Пьером Гиро [50] для французского языка:

$$S=C_{00}\sqrt{f}$$

где S - количество значений у лексемы, f – частота употребления соответствующего слова в речи, C_{00} – коэффициент пропорциональности.

Кстати, другой лингвист (Мак Вей) установил, что для чешского языка частота (f) должна стоять под корнем не второй, четвёртой степени.

Однако сама идея этой закономерности чрезвычайно ценна. Она состоит в том, что многозначность лексемы (S) зависит от её частотности (f).

Дело в том, что ни один лингвист рационально не объяснил причину появления многозначных (полисемичных) слов (точнее, лексем). Ведь известно, что все слова возникают как единство звучания и одного (!) значения.

Однако вспомним, что К. Ципф [68] установил, что в английском языке существуют две следующие закономерности: «*высокочастоные слова фонетически короче и семантически многозначнее, чем низкочастотные!*»

Если использовать более крупный и рациональнее составленный частотный словарь английского языка, в котором указана частотность каждого из 50 тыс. слов (объём выборки составляет 1 млн. СУ) **[56],** то оказывается, что эмпирические формулы, описывающие частотные слова французского языка, вполне пригодны для описания частотных слов английского языка!

Используем философскую категорию «аналогии», в частности,
применим эмпирические формулы, придуманные Пьером Гиро [50] для описания французского языка, к описанию закономерностей английского языка, обнаруженных К. Ципфом.

Итак, закон Ципфа-Гиро [50; 68], названный нами фонетическим:

$$L= C_1 / \lg f$$

где L –количество фонем в лексеме, f – частотность употребления слова, C_1 – фонетический коэффициент пропорциональности (для русского, английского языков $C_1 \approx - 40$). Закон формулируется следующим образом.

«Фонетическая длина (L) лексемы обратно пропорциональна частотности (f) употребления соответствующего слова в речи», то есть, «Чем выше частотность слова, тем оно стремится стать короче».

Этот закон объясняет одну из причин того, почему, например, во французском языке «длинные» слова латинского происхождения «укоротились»).

По аналогичной причине множество английских слов «укоротилось» за время, прошедшее с момента их первой письменной фиксации (заметим, что в хронологическом Оксфордском словаре [63] самые древние слова датируются VIII веком наше эры).

Вводим следующий закон Ципфа-Гиро, названный нами **семантическим:**

$$S=C_2 \sqrt{f}$$

где S - количество значений у лексемы, f – частота употребления соответствующего слова в речи, C_2 - семантический коэффициент пропорциональности (подсчитано, что для русского и английского языков $C_2 \approx 1000$). Его можно сформулировать следующим образом.

«Количество значений (S) у лексемы прямо пропорционально частотности (f) употребления соответствующего слова в речи», то есть «Чем чаще слово употребляется в речи, чем более многозначной (полисемичной) стремится стать соответствующая лексема».

Иными словами, одной из объективных причин появления многозначности (полисемичности) является высокая частотность слова в речи.

. Заметим, что ни в одной работе по семантике или семасиологии вопрос о причинах полисемии даже не ставится, так как авторы не знают на него ответа.

Интересно, что, хотя П. Гиро [50] и установил, что высокочастотные французские слова являются древними, но формулу, описывающую эту закономерность, он не приводит. Учитывая тот факт, что французский и английский языки относятся к родственным группам индоевропейской семьи (кстати, в английском языке около 60% - 70% слов, заимствованных из французского языка) мы воспользуемся философской **категорией «аналогии»** и предложим следующий вид этой зависимости, которую из уважения к первооткрывателям назовём хронологическим законом Ципфа-Гиро.

$$T=C_3\sqrt{f}$$

где T – возраст лексемы, f - частотность употребления слова в речи,

C_3 - хронологический коэффициент пропорциональности.

«Возраст (T) лексемы прямо пропорционален частотности (f) соответствующего слова в речи», т.е. «Чем выше частотность слова, тем оно старше (древнее)».

Таким образом, эти три закона описывают связь основных характеристик слов и лексем английского языка (звучание, значение и возраст) с частотностью употребления слов в речи, формируемой обществом, то есть связь между «внутренней» и «внешней» лингвистикой (в частности, социолингвистикой).

Обратим внимание на то, что аргумент во всех этих формулах один и тот же: f - частотность. Объединив попарно эти уравнения, можно решить получившиеся «системы из двух уравнений с одним неизвестным» в общем виде и получить три следствия из законов Ципфа-Гиро:

Фоносемантическое следствие из законов Ципфа-Гиро (в одной форме).

$$L= C_4 / S$$

«Фонетическая длина лексемы обратно пропорциональна ее многозначности (полисемичности)».

Семантикофонетическое следствие из законов Ципфа-Гиро (в другой форме):

$$S= C_4 / L$$

«Количество значений у лексемы обратно пропорционально её фонетической длине»»

Семантико-хронологическое следствие из законов Ципфа-Гиро (в одной форме).

$$S=C_5T$$

«Количество значений у лексемы прямо пропорционально ее возрасту».

Отметим, что этот закон объясняет вторую причину многозначности (полисемичности) лексем, а именно: «Чем старше лексема, тем она многозначнее (полисемичнее)».

Хронологосемантическое следствие из законов Ципфа-Гиро (в другой форме):

$$T=S / C_5$$

«Возраст лексемы прямо пропорционален ее многозначности (полисемичности)».

Фонетико-хронологическое следствие из законов Ципфа-Гиро (в одной форме):

$$L=C_{6\,/}\,T$$

«Фонетическая длина лексемы обратно пропорциональна ее возраст»

Хронофонетическое следствие из законов Ципфа-Гиро (в другой форме).

$$T=C_{6\,/}\,L$$

«Возраст лексемы обратно пропорционален ее фонетической длине»

К концу XX века появился ряд работ, авторы которых подсчитали количественные величины ряда характеристик лексем английского языка.

Так, группа преподавателей Ленинградского университета под руководством и при деятельном участии профессора И. Ивановой качественно-количественно описала суффиксы английских имен существительных, содержащиеся в Кратком Оксфордском словаре среднего объёма Было подсчитано количество разных слов с каждым из 75 суффиксов существительных, приведены словообразовательные значения суффиксов с примерами, определена частеречная принадлежность ПО, подсчитана морфемная членимость дериватов, подсчитаны доли конечных фонем ПО (доли которых статистически не обрабатывались).

По точно такой же методике и по тому же источнику группа преподавателей Дальневосточного университета описана 27 суффиксов прилагательных а позже и суффиксы наречий и числительных английского языка Таким образом, вся суффиксальная словообразовательная подсистема английского языка была описана качественно-количественно.

Как известно, ещё в 20-х годах XX века Торндайк [65] начал составлять частотные словари английского языка**).**

В 1934 году вышел четырёхязычный словарь Итон [48], содержащий список тысячу самых частотных слов английского, французского, немецкого и испанского языков, который можно использовать для рационального и быстрого изучения лексики этих языков.

Первый частотный словарь русского языка вышел в 1953 году в Америке [54]. Он содержал всего 5 000 слов (объём выборки был равен 1 млн. словоупотреблений).

И, наконец, Частотный словарь русского языка [47], содержащий 40 000 слов с указанием частотности каждого слова (выборка исходного материала равнялась 1 млн. словоупотреблений, что в наше время считается достаточной для получения надежных результатов.)

В 1967 году вышел частотный словарь современного английского языка, содержащий 50 тыс. слов с указанием частотности каждого слова (объем выборки равен 1 млн. словоупотреблений [56].

Ещё в 1980 году сообщалось, что «к настоящему времени составлено и опубликовано около 500 частотных словарей более 30 языков» [45, 266].

Далее говорится, что «в нашей стране составлено около 150 отраслевых частотных словарей, и почти все они созданы общесоюзной группой «Статистика речи», руководит которой проф. Р.Г. Пиотровский» [45, 267].

Одним из первых в СССР появился частотный англо-русский словарь-минимум по электронике [3], причем объём выборки равнялся всего 200 000 словоупотреблений (кстати, эта работа послужила основой для кандидатской диссертации автора. (Кстати, ВАК вскоре запретил принимать к защите диссертации, основой которых было составление отраслевого частотного словаря и описание некоторых его статистических характеристик. То есть «эту лавочку прикрыли»).

Позднее к этому словарю добавили ещё два (по физике твёрдого тела и физике элементарных частиц), и объём выборки теперь составляет 600 000 СУ, а словарь содержит 4 666 слов с частотностью 2 и более (а жаль, что были исключены слова с частотностью равной единице: их было около 3 тысяч, следовательно, значительную часть информации о физических терминах авторы «недобрали». Ну да физики им судьи!).

В словаре М. Уэста [67] содержится 2000 высокочастотных английских слов и указывается частотность употребления каждого значения, приведенного в БОС (1933). Это был единственный частотностей значений слов на тот момент.

А.Н. Тихонов [35], составивший Словообразовательный словарь русского языка объёмом в 145 тыс. слов [36] (крупнейший на тот момент!), проделал огромную работу по количественной его характеристике. Оказывается, 18,8 тыс. слов являются непроизводными, а остальные можно объединить гнёзда: существительные входят в 8,9 тыс. гнезд (66,7 тыс. слов), глаголы – в 2,4 тыс. гнёзд (56 тыс. слов.)), прилагательные – в 944 гнезда (14, 6 тыс. слов), наречия – 85 гнёзд (716 слов) и т.д., что впервые стало известно русистам. Интересно, что количество гнездовых слов (138 576 штук) в 25 раз превосходят удельный вес одиночных (5 576 штук). Извлекши величины частотности 20 высоко частотных и 20 низкочастотных, он установил, что словообразовательный потенциал высоко частотных глаголов (152 производных на слово), чем у низко частотных (19 производных га слово). Это означает, что высокочастотные глаголы образуют и имеют больше производных слов. Аналогичная ситуация и с существительными и прилагательными.

Анализ словообразовательных потенциалов многозначных слов показал, что у глаголов он равен 140 производных на слово, а у однозначных – 8,6; у существительных соответственно – 82 и 7,6; у прилагательных – 81 и 4,7 производных на слово. Это означает, что многозначные слова дают в 16 – 17 раз больше производных, чем однозначные [35].

В современной лингвистике категория количества может выступать в виде коэффициента корреляции, в частности, так называемых тетрахорических коэффициентов связи между качественными признаками. Их применяют тогда, когда каждый элемент 2-х описываемых совокупностей двумя наборами величин характеризуются двумя свойствами (качествами).

Профессор Ю.А. Тулдава [38], изучавший частотные словари, вывел формулу для нахождения количества разных слов (лексем) (L) в частотном словаре в зависимости от объема использованной выборки (V)^

$$L = a\ V \exp b,$$

где a и b – параметры

Кроме того, он предложил использовать формулу Чекановского [49] (правда не сослался на этого автора!) для нахождения доли (R) общей лексики при объединении 2-х частотных словарей в одно целое

$$: R = 2\ C / (A + B),$$

где A и B – объёмы сравниваемых словарей, C – объём общей (совпадающей) лексики.

Ценным является наблюдение профессора Ю.Тулдавы [38, 18], что «в лингвистической теории и практике по разным основаниям…выделяются парные (категории): социальное – индивидуальное, идеальное – материальное, общее – отдельное, абстрактное – конкретное и др., которые, отдельно взятые или в комбинации, служат для определения и интерпретации понятий «языка» и «речи» [38, 18].

потенция
(язык»)
динамика________________|________________ статика
|
реализация
(«речь»)

Далее профессор Ю. Тулдава [38, 18] графически изображает важное соотношение между двумя парами лингвистических категорий и поясняет: «Противопоставление *потенции* и *реализации* рассматривается в данном случае как собственно языковое-речевое соотношение, то есть, «языком» считается система потенциальных возможностей, а «речью» - актуализация этих возможностей в действительности» [38, 18].

Одной из первых книг по «математической лингвистике» вышла в СССР в 1961 году [6].

Одна из авторов - Р. М. Фрумкина – ратует за «применение статистических методов в изучении языка» [6, 67]. Она разъясняет народу «понятие частоты, выборки и относительной ошибки» [6, 67]. и приводит важную формулу для нахождения объёма выборки (N) [6, 67] использующуюся в теории вероятностей [20].

$$N = t\,t\,q\,/\,d\,d\,p,$$

где t – критерий Стьюдента, d – относительная ошибка, p - доля интересующего нас элемента, q – доли остальных элементов выборки.

Она также приводит формулу Чотлоса [61] для подсчёта количества разных слов (L) в выборке объёмом в N

$$L = N (a - \ln N) / b,$$

где a, b – константы, находимые эмпирически.

Авторесса пишет, что «в эксперименте Чотлоса при a = 11,670 и b = 11,268 эта формула хорошо описывала кривую прироста новых слов для N < 18 000; Однако проверка этой формулы показывает, что при N > 18 000 она неудовлетворительна» [6, 88].

Далее в работе упоминается так называемая формула Ципфа (которая на поверку оказалась тавтологическим соотношением, так как величины P и r практически линейно связаны).

$$P\ r \exp(-y) = k,$$

где r – номер слова в ранжированном по убывающей частотности ряду, P – его вероятность, k и y- константы.

Но критики говорят: «Представим себе взвод солдат (это, скажем, 20 человек), выстроенных «по росту», то есть от максимального к минимальному, и скомандуем: «По порядку номеров рассчитайсь!». Это значит, что самый рослый солдат получит №1, следующий - №2 и т. д., а самый низкорослый получит самый большой номер (№20). Вот подобную ситуацию и описывает эта так называемая формула Ципфа, хотя любому видно «невооружённым глазом», что величина роста солдата постепенно уменьшается к концу ряда, а его порядковый номер почти «пропорционально» увеличивается.

Однако, есть и некоторая польза от этого выражения: если взять k = 0,1, а y = 1,01, то можно показать, что, например, сумма долей слов от 1-го до 1100 оказывается равной 0,8 (это расшифровывается следующим образом: если ученик выучит 1100 самых частотных слов, то он будет понимать (то есть переводить) 80% текста – а это не мало!).

В 1971 году вышла книга профессора Б.Н. Головина [17]. посвященная статистическому описанию лексики русского языка. Автор находит средние значения величины ряда характеристик, приводит и использует в своей работе статистические формулы для определения средних значений, среднего квадратического, критерия Стьюдента, Хи-квадрат, коэффициента вариации, формулу для определения достоверности получаемых величин.

Следует отметить, что эта книга написана так, что её могут понять и использовать лингвисты, «не имеющие высшего образования по математике», чего нельзя сказать о книге Носенко [25].

Начали появляться обобщающие работы по математической лингвистике [6; 30].

Параллельно идут попытки по использованию формул, извлекаемых из бездонного «рога изобилия» математики, к описанию различных закономерностей, которым подчиняются лингвистические элементы или их множества.

Использование методов математической статистики, начавшееся в начале XX века, продолжается и поныне. Лингвистам приходится вникать во все «математические тайны», постигая такие понятия, как объём выборки, среднее квадратическое отклонение, уровень значимости, коэффициент вариации, критерий Стьюдента, дисперсия – момент второго порядка, асимметрия – момент третьего порядка, эксцесс – момент четвёртого порядка, корреляция линейная, криволинейная, множественная; коэффициенты корреляции ранговые (Спирмена и Кендела) и тетрахорические (Чекановского, Жаккара и др.), сравнение 2-х совокупностей, информционный коэффициент Шеннона и прочая математическая абракадабра.

Взаимодействие лингвистов с математиками порой осуществляется, как в известном анекдоте.

«Встречает лингвист на улице своего друга математика и видит, что у него на ногах разные ботинки (ну, там по цвету, фасону, размеру – не важно). Он и говорит математику; «Слушай, у тебя на ногах разные ботинки». Тот посмотрел и сокрушенно говорит: «Действительно…Что же делать?». Ну, лингвист ему и говорит: «Пойди домой и переобуйся»! Математик подумал-подумал и говорит: «А что я найду дома? Там тоже два разных ботинка». Мораль: вот так всегда: лингвист ставит математикам задачу, а они не знают, как её решить!

Как известно, «Корреляционный анализ сводится к измерению тесноты или степени сопряженности между варьирующими признаками, а также к определению формы и направления существующей между ними связи» [20, 171].

«Наряду с параметрическими применяются и непараметрические, или [порядковые критерии» [20, 111; 31-33].

Научная школа профессора Г.Г. Сильницкого [31-33] и ныне уже профессора С.Н. Андреева [5] исследует английские глаголы из The Concise Oxford Dictionary [63]. Путем сплошной выборки было извлечено 8279 глаголов и подсчитаны корреляционные связи между 70 качественными признаками каждого глагол, используя ЭВМ. Было отобрано с полдюжины коэффициентов корреляции различных авторов (для повышения надёжности получаемых результатов): Пирсона, Юла, Коула, Михаеля, Хильберта (которые имели пределы изменения от +1 до –1) [18]; Также использовались коэффициенты сходства Брея, Иверсена, Дайса-Соренсона, Жаккарда, Рассела, Миченера, которые могут принимать значение от 0 до +1 [18].

Члены научной школы профессора Г.Г. Сильницкого на огромном экспериментальном материале оценили степень точности «работы» разных коэффициентов, что является ценным вкладом математическую лингвистику.

Наиболее известным непараметрическим показателем связи является ранговый коэффициент корреляции Спирмена [60], определяемый по формуле

R = 1 – 6 Sum d*d / n (n+1) (n-1) (-1 < R < +1)

Здесь: * - знак умножения, Sum – знак суммирования; d – разность между рангами сопряженных значений признаков X и Y, т. е. d = Sum (Xi – Yi) exp 2; где n – объём выборки или общее число парных наблюдений; d*d – это d в квадрате» [20, 171].

Для оценки достоверности полученной величины коэффициента ранговой корреляции можно воспользоваться так называемым критерием Стьюдента

t = r / exp 0,5 [(t – r*r) / (n - 2)]

Из теории известно, что при 95% достоверности полученной величины r значение t должно быть меньше 1,96 (то есть t < 2). Проиллюстрируем это.

Во многих языках, в том числе и в английском, наблюдается явление параллелизма, то есть, например, образование одноосновных разносуффиксных дериватов с суффиксами прилагательных -ish и –ly. Подсчитаем величину рангового коэффициента **корреляции** r по полученным нами данным (Тал. 2).

Таблица 2. Подсчёт величины рангового коэффициента корреляции Спирмена

	-ish	-ish	-ly	-ly		
Производящ. основы	Дата возникн.	Xi	Дата возникн.	Yi	D (i)	D (i) в квадрате
Fiend-	1529	7	1050	9	--2	4
Goat-	1528	6	1850	2	4	16
Maiden-	1749	4	1450	6	-2	4
Man-	888	9	1200	8	1	1
Pig-	1792	2	1859	1	1	1

Tiger-	1573	5	1633	5	0	0
Quaker-	1787	3	1696	3	0	0
Woman-	1390	8	1374	7	1	1
Vixen-	1828	1	1677	4	3	9
Сумма	-	-	-	-	-	36

Ранги Xi и Yi представляют собой номера дат первой письменной фиксации дериватов в порядке убывания, начиная с самой большой величины. Найдём сумму квадратов разностей рангов и подставим в формулу:

$$r = 1 – 6 *36 / 9 * 80 = 0,70$$

Для определения достоверности величины r, подсчитаем величину критерия

Стьюдента t.t = 0,70 / exp 0,5 (1 – 0,49) / 7 = 2,3

Итак, получаем величину t = 2,3. Это больше 2-х, следовательно, величина r достоверна!

Известен также ранговый коэффициент корреляции Кендэла Т (тау)

$$T = 1 – 4 Q / n (n-1) \qquad (-1 < T < +1)$$

Продемонстрируем нахождение величины коэффициента корреляции Кендэла (Таблица 3).

Для подсчёта величины Т необходимо ранжировать конверсификсы по одному из критериев, например, по Ps. Затем необходимо найти число «совпадений» или «инверсий» рангов обоих критериев. «Совпадения» подсчитываются следующим образом. Так, конверсификс –up по критерию Pd имеет ранг, который по величине меньше всех 8-ми ниже расположенных конверсификсов. У конверсификса –out ранг меньше всех 7-ми ниже расположенных морфем. У конверсификса -in ранг меньше 5-ти расположенных ниже морфем и т. д. Сумма всех этих величин даст нам сумму Р («совпадений»), которая фигурирует в формуле. Для нахождения количества «инверсий» необходимо подсчитать, в скольких случаях ранг конверсификса –

up больше рангов всех ниже расположенных морфем, и т. д. Сумма всех этих величин даст Q (Табл. 3). Подсчёт величины коэффициента Кендэла можно производить по приведённой формуле (знак * означает умножение).

$$T = 1 – 4*4 / 9*8 = 0,778$$

Для быстрого нахождения приблизительной величины силы связи часто используется коэффициент Фехнера [40].

$$I = (v-w) / (v+w)$$

Где v – количество совпадений, а w – количество несовпадений знаков разностей между абсолютными значениями критериев каждого конверсификса и соответствующими средними значениями (Таблица 3). Подсчитаем число совпадений (- , +) в столбцах 4 и 5. Получим v = 7. Соответственно количество несовпадений: w = 2. Подставив эти величины в формулу, получим величину I.

I = (7 - 2) / (7 + 2) = 0,56.

Таблица 3. Подсчёт величин коэффициентов корреляции Спирмена и Фехнера для оценки силы корреляции между параметрами Ps и Pd конверсификсов английского языка

Конверсификсы	Ранг Ps	Ранг Pd	Совпадения	Инверсии	Коэффициенты
-up	1	1	8	0	
-out	2	2	7	0	
-in	3	4	5	1	
-down	4	5	4	1	
-off	5	3	4	0	
-on	6	6	3	0	
-over	7	9	0	2	
-away	8	7	1	0	
-by	9	8	0	0	
Коэффициенты корреляции					R=0,900 T=0,778

В последнее время в лингвистике для оценки силы связи между различными характеристиками аффиксов и производных единиц используют различные коэффициенты связи (корреляции, ассоциации) [5; 31-33]. Известно, что такие коэффициенты используются в биологии, геоботанике, экологии, педагогике и психологии [18]. Чаще всего применяется коэффициент Пирсона [18]:

$$\Phi = (ad - bc) / \exp 0{,}5\,(a + b)\,(a+c)\,(b+c)\,(c+d) \qquad (-1 < \Phi < 1)$$

Ф – коэффицент Пирсона, a, b, c, d – это количество случаев наличия или отсутствия каждого из двух признаков.

Заметим, что хотя этот коэффициент имеет удобные пределы изменения, его Величина Ф резко нелинейно зависит от величин a, b, c, d, что затрудняет интерпретацию полученных величин (результатов). Кроме того, его нахождение требует довольно громоздких вычислений, на что указывают некоторые специалисты по статистике [40].

Относительно широко распространён коэффициент ассоциации Юла, составленный из тех же самых величин четырехпольной таблицы.

$$Q = (ad - bc) / (ad + bc) \qquad (-1 < Q < 1)$$

Заметим, что этот коэффициент тоже нелинеен относительно величин a, b, c, d, но вычислять его значительно проще, чем коэффициент Пирсона.

Существует несколько более просто устроенных тетрахорических коэффициентов корреляции:

К ним относится менее известный коэффициент Гамана [51], имеющий вид

$$G = (a-b-c+d) / (a+b+c+d) \qquad (-1 < G < +1)$$

Отметим, что он прост по своему устройству и имеет удобные пределы изменения (типичные для коэффициентов корреляции).

Также несложен в употреблении коэффициент Рассела:

$$Kr = a / (a + b + c + d) \qquad (\mathbf{0} < Kr < \mathbf{1})$$

Не менее просто устроен и коэффициент Миченера

Km = (a+d) / (a+b+c+d)   **(0 < Km < 1)**

Заметим, что все эти коэффициенты получены путём более или менее удачного манипулирования характеристиками элементов a, b, c, d.

В ряде случаев используются коэффициенты сравнения 2-х совокупностей несложного вида.

Кэффициент Чекановского [49]:

Kc = 2c / (a+ b)   (0 < Kc < 1)

Коэффициент Жаккара [52]

Kz = c / (a+b-c)   (0 < Kz < 1)

Коэффициент Брея

Kbr = 2a / (a+b+c)   (0 < Kbr < 1)

Коэффициент Дайса-Соренсена

Kds = 2a / (2a+b+c)   (0 < Kds < 1)

Коэффициент Иверсена

Ki = a / (a+b+c)   (0 < Ki < 1)

Но всех коэффициентов сравнения только коэффициент Чекановского научно обоснован. Проиллюстрируем это следующим рассуждением.

Пусть мы сравниваем 2 множества разных объектов с мощностью «а» и «b». Мощность множества элементов, одинаковых для обоих множеств, будет «с». Как известно, при пересечении 2-х множеств пересекающаяся части накладываются друг на друга, следовательно, в пересекающейся части мы имеем двойной набор элементов: 2с. Отсюда и формула:

Kc = 2c / (a + b)   (0 < Kc < 1)

В отличие от всех прочих этот коэффициент изменяется пропорционально, то есть если «с» увеличится в 2 раза, то и Кс возрастёт в 2 раза. Если «с» увеличится в 5 раз, то и Кс возрастёт в 5 раз.

Авторы всех прочих коэффициентов просто жонглировали буквами, комбинируя их «как богу (или чёрту) угодно», беря их просто «с потолка» или «наобум лазаря» абы лишь бы «попасть в историю» (как говаривали украинские бандеровцы во времена оны).

В свое время появилось обобщение коэффициента Жаккара на n-мерный случай, то есть пригодного для сравнения n-совокупностей:

$$Kk = (n-1)\, m / (S1+S2+S3,\ldots, +Sn - m), \qquad (0 < Kk < 1)$$

n – количество сравниваемых совокупностей, m – количество элементов, одинаковых для всех совокупностей, S1, S2, S3, …, Sn – количество элементов в n совокупностях. К сожалению, этот коэффициент (Kk) тоже нелинеен, как и коэффициент Жаккара (обобщением которого он является.).

Проанализировав ситуацию, мы сконструировали расширение коэффициента Чекановского Kc на n-мерный случай. Это коэффициент Барткова [12]

$$Kб = n\, m / (S1+S3+\ldots+Sn), \qquad (0 < Kб < 1),$$

где n – количество сравниваемых совокупностей, m – количество одинаковых элементов во всех выборках, Si, S2, ,,,, Sn – количество элементов в 1-й, 2-й, 3-й и т.д. выборках.

Отметим, что наш коэффициент сравнения «пропорционален» изменениям величины m, то есть линеен.

В 1952 году американский лингвист М. Сводеш [62] предложил метод для количественного определения времени, прошедшего с момента распада 2-х родственных языков, названный методом глоттохронологии, или лексико-статистики. В то время уже было известно, что в каждом языке со временем исчезает (забываются) определенное количество слов. Он предложил составить список из 200 слов, обозначающих самые общие понятия (например: большой, птица, грудь, ноготь, пить, глаза, ухо и т.д.), а затем считать доли слов из этого списка, сохранившихся в каждом из двух родственных языков за тысячу лет. Чем меньше доля сохранившихся в обоих языках слов «С», тем больше времени «t» прошло с момента их разделения. В том же 1952 году другой американский лингвист предложил формулу для подсчета времени, прошедшего с момента разделения языков:

$$t = \log r * r / \log C$$

где r – индекс сохранения (доля) основного списка за одно тысячелетие, который равен в среднем (для обследованных на тот момент языков) 0,85.

Величина С означает долю совпадающих слов в двух языках.

Наиболее широкое применение этот метод нашел при изучении индейских языков Америки, которые не имели письменности, поэтому невозможно было использовать сравнительно-историческим метод.

Только применение глоттохронологического метода позволило разобраться в очень сложной картине родства языков американских индейцев (как известно, предками индейцев были племена северо-восточных азиатских охотников, которые двумя волнами – 20 тыс. и 35-40 тыс. назад – прошли по Берингийской суше на Аляску и хлынули далее на юг и восток Северной Америки, причём некоторые племена из первой волны достигли в конце концов южной оконечности Американского континента, например, патагонцы). Если первая волна двигалась по незаселённой местности на юг, восток и запад, через Панамский перешеек и далее на юг, то второй волне пришлось «с боями» продвигаться, отыскивая себе «жизненное пространство»! Поэтому почти все они разместились в Северной Америке и на перешейке между местообитаниями ранее прибывших в составе первой волны. Все это привело в тому, что картина размещения племён оказалась очень пестрой.

Один американский исследователь финно-угорских языков, используя метод глоттохронологии, установил, что ненецкий язык отделился от уральского предка шесть тысяч лет назад, что совпало с результатом, полученным другими методами ранее.

Используя метод глоттохронологии, было показано, что грузинский и занский языки начали расходиться около VII в. до н.э., грузинско-занский и сванский языки начали расходиться ещё в XIX в. до н.э.

Интересно, что пара лингвистов (очевидно, бывших троечников по математике) Л.С. Бархударов и Г.В. Колшанский – огульно охаяли метод глоттохронологии, заявив, что «математический анализ накладывается здесь как совершенно чуждый природе языка метод». Кстати, Л.С. Бархударов – это «неразумное чадо» известного советского лексикографа С. Бархударова!

Но они «сели в лингвистическую лужу» и «попали пальцем небо» из-за своего невежества, а этот метод успешно применяется и поныне, помогая решать многие проблемы родства языков, которые невозможно решить другими методами.

Как известно, предложенная в 1948 году К. Шенноном [58] формула для оценки количества информации, проходящей по каналу связи, быстро революционизировала проблему использования математики при описании самых различных систем. Появилась работы, посвященные её использованию в лингвистике [29].

Мы используем её в качестве коэффициента неоднородности распределения долей.

Если в каждом классе будут находиться одинаковые доли элементов, то

$$H^* = 1 - \text{Sum } Pi \log Pi / \log n,$$

где H* - относительный коэффициент неоднородности распределения долей элементов по классам, Pi – доля элементов i-го класса, log Pi – двоичный логарифм доли Pi, причём номера классов «пробегают значения» i=1,2,…, n; Sum – знак суммирования произведений Pi log Pi. Если К=1, то неоднородность распределения элементов по классам максимальна (все элементы в одном классе), если К=0, то неоднородность минимальна (налицо однородность) (то есть элементы поровну распределяются по классам).

Словообразовательные гнёзда давно интересуют лингвистов. Они формируются в языке в течение длительных промежутков времени, накапливая производные, образованные различными способами. Особенно многочисленные гнёзда содержат как простые слова, так и аффиксальные производные. Представляет интерес оценить соотношение между префиксальными, суффиксальными дериватами, а также между препозитивными и постпозитивными корнями по отношению к ядерному корню.

Мы подсчитали количество членов гнёзд, образованных разными способами от 16 разных ядерных лексем с целью установления степени их неоднородности, а для объективной оценки воспользовались информационным коэффициентом H* [68].

Таблица 4. Подсчёт коэффициентов неоднородности (H*)

17 словообразовательных гнёзд

Ядро	Кол-во	P1	P2	P3	P4	H*	Кс
Man,n	472	0,01	0,96	0,02	0,01	0,9244	0,9217
Land,n	267	-	0,60	0,38	0,02	0,7286	0,5048
House,n	236	0,01	0,64	0,34	0,01	0,7315	0,5254
Water,n	225	0,01	0,26	0,71	0,02	0,7212	0,5722
Stone,n	207	-	0,58	0,40	0,02	0,7256	0, 4968
Wood,n	167	0,01	0,60	0,34	0,05	0,6865	0,4782
Berry,n	71	-	0,96	-	0,02	0,9648	0,9608
Monger,n	22	-	0,92	-	0,05	0,9284	0,9050
Smith,n	17	-	0,65	0,11	0,24	0,6879	0,4922
Wright,n	6	-	1,00	-	-	1,0000	1,0000
Like,a	124	0,03	0,90	0,01	0,06	0,8504	0,8146

Grey,a	24	-	0,50	0,33	0,17	0,6344	0,3878
Well,adv	1173	0,001	-	0,99	-	0,9971	0,9980
Ill,adv	210	-	-	0,98	0,02	0,9646	0,9608
Say,v	8	0,38	0,24	-	0,38	0,6113	0,3454
Sit,v	7	О,28	0,44	-	0,28	0,6126	0,3504

Примечание. В последнем столбце приведены величины коэффициента Симпсона (Кс), который рассмотрим ниже. Классы имеют следующие обозначения: Р1 – префиксальные, Р2 - препозитивные корни по отношению к ядру, Р3 – постпозитивные корни сложений, Р4 – суффиксальные.

Гнёзда, у которых все производные образованы одним способом (то есть у них Р1=1,00), имеют Н*=1,00. Например, в гнезде существительного Wright

Здесь все производные – это сложные слова с препозитивными существительными.

Гнёзда с ядерными существительными Man, Berry, Monger почти полностью состоят из препозитивных существительных плюс ядро, то есть очень неоднородны. На это указывают величины их Н* (от 0,93 до 0, 96).

У прилагательного Like гнездо на 90% состоит из прилагательных с именным первым компонентом, поэтому у него высокое значение Н*=0,8504.

У наречных гнёзд с ядрами Well, Ill гнёзда на 98%-99% состоят из сложений, стоящих в постпозиции к ядру, поэтому у них очень высокие величины Н* (0,96-0,99).

Интересно, что в глагольных гнёздах с ядрами Say, Sit дериваты разных моделей образования распределены по трём классам примерно равномерно (в пределах от 0,28 до 0,44). На это указывают величины К*=0,61.

Существует ещё один коэффициент для оценки неоднородности распределения элементов по классам – вероятностный, предложенный Симпсоном [59].

Kci = Sum Pi*Pi (i = 1, 2, …, n)

где Pi – это доля элементов, попавших в i-тый класс, Pi*Pi – это Pi в квадрате, n - количество классов, Sum – это сумма квадратов долей классов.

Необходимо заметить, что пределы изменения этого коэффициенты зависят от величины n – количества классов, что неудобно при сравнении совокупностей, имеющих разное количество классов. Здесь если Kci=1,00, то распределение по классам резко неоднородно (например, когда все элементы попадают в один класс!); если Кс=1/n, то распределение максимально равномерно.

Гнёзда, у которых все производные образованы одним способом (то есть у них P1=1,00), имеют H*=1,00. Например, в гнезде существительного Wright Все производные – это сложные слова с препозитивными существительными. Гнёзда с ядерными существительными Man, Berry, Monger почти полностью состоят из препозитивных существительных плюс ядро, то есть очень неоднородны. На это указывают величины их H* (от 0,93 до 0, 96).

У прилагательного Like гнездо на 90% состоит из прилагательных с именным первым компонентом, поэтому у него высокое значение H*=0,8504.

У наречных гнёзд с ядрами Well, Ill гнёзда на 98% - 99% состоят из сложений, стоящих в постпозиции к ядру, поэтому у них очень высокие величины H* (0,96-0,99).

Интересно, что в глагольных гнёздах с ядрами Say, Sit дериваты разных моделей образования распределены по трём классам примерно равномерно (в пределах от 0,28 до 0,44). На это указывают величины К*=0,61.

Возраст лексем – это одна из объективных причин возникновения полисемии.

Для решения проблемы полисемии важным является следствие из законов Ципфа-Гиро, гласящее: «Чем старше лексема, тем она многозначнее», т.е. одной из объективных причин полисемии является возраст лексемы.

Еще в 1982 году, диахронически описывая высокочастотные дериваты с глагольным суффиксом –ify в английском языке, мы [10; 13] впервые обнаружили следующую эмпирическую закономерность: лексемы, возникшие в XIII в., имеют в среднем по 7,0 значений на лексему (то есть, S*=7,0); у лексем, возникших в XIV в., S*=6,0 знач/лекс; в XV в. - S*= 6,5 знач/лекс; в XVI в. – S*=5,7 знач/лекс; в XVII в. - S*=2,0 знач/лекс; в XVIII в. – S*=1,0 знач/лекс; в XIX в. – S*=2,0 знач/лекс (Табл. 5).

Таблица 5.Изменение средней полисемичности (S*) английских глаголов с суффиксом -ify во времени

Век	Fm	S	S*
XIII	1	7	7,0
XIV	9	54	6,0
XV	4	26	6,5
XVI	3	17	5,7
XVII	1	2	2,0
XVIII	1	1	1,0
XIX	1	2	2,0
XX	-	-	-
Сумма	20	101	5,1
Среднее			

Примечание: Среднюю полисемичность лексем (S*) подсчитывали по формуле: S*= S / Fm , где Fm - количество высокочастотных слов, возникших в данном веке; S - суммарное количество значений у этих лексем.

Анализ величин S* показал, что существует следующая зависимость: «Чем старше лексема, тем она многозначнее (полисемичнее)» Это явилось первым прямым экспериментальным объяснением существования явления многозначности английских лексем.

Приведем еще один пример зависимости полисемичности лексем от их возраста (Табл. 6). Так, было экспериментально установлено, что в уникальном диахроническом словаре английского языка [64] объемом в 616 тыс. статей содержится 1316 лексем с заимствованным суффиксом –ism. Лексемы с этим суффиксом стали проникать в английский в XIV веке. Анализ их семантики показал, что наиболее полисемичными являются лексемы, возникшие в XIV и XV вв.: S*=4,4 знач/лекс и S*=3,3 знач/лекс соответственно. Лексемы, возникшие в начале новоанглийского периода, в среднем имеют S*(XVI в.)=2,5 знач/лекс и S*(XVII в.)=2,1 знач/лекс. Лексемы, появившиеся в поздний новоанглийский период, характеризуются следующей средней полисемичностью (S*): S*(XVIII в.)=1,8 знач/ лекс; S*(XIX в.)=2,1 знач/лекс; S*(XX в.)=1,1 знач/лекс.

Следовательно, независимо от того, сколько лексем появилось в тот или иной век, четко прослеживается общая зависимость между возрастом и полисемичностью лексем: «Чем старше лексема, тем она многозначнее (полисемичнее), и наоборот» (Табл.6).

Таблица 6. Изменение средней полисемичности дериватов с суффиксом –ism в диахронии

Век	Ps	S	S*	Примеры
XIV	5	22	4,4	Embolism, exorcism
XV	3	10	3,3	Graecism, Judaism
XVI	39	97	2,5	Gentilism, Puritanism
XVII	156	320	2,1	Nepotism, stoicism
XVIII	121	220	1,8	Heroism, Sabianism

XIX	815	1161	1,4	Industrialism, nominalisn
XX	177	199	1,1	Gaullism, lettrism
Сумма	1316	2029	-	
Среднее	-	-	1,5	

Примечание: Средняя полисемичность лексемы (S*) подсчитывали по формуле: S*= S / Ps, где Ps - количество лексем, появившихся в течение века;
S - сумма значений всех этих лексем.

Позднее эта закономерность многократно подтверждалась нами при полном количественном диахроническом описании каждого из 30 английских префиксальных и суффиксальных деривационных типов.

Заметим, что в 1977 году Р.А. Будагов в книге «Что такое развитие и совершенствование языка?” [15] написал об открытии им «закона многозначности слов», сообщив, что если мы «раскроем большой Толковый словарь русского или польского, французского или испанского, английского или немецкого языков и подсчитаем, сколько многозначных слов находится на той или иной странице, то получится, что около восьмидесяти процентов слов… предстанут перед нами как слова многозначные».

Отметим, во-первых, что «слова», как единицы речи, всегда однозначны, а «лексемы», которые мы находим в толковых словарях, в начале своей «лингвистической карьеры» всегда однозначны, но со временем, часто употребляясь в речи, могут стать многозначными. Известно, что в языках цивилизованных народов насчитывается около 300 – 600 тыс. лексем (например, в «Новом большом русско-английском словаре» [24] содержится более 300 000 лексических единиц русского языка; в The Oxford English Dictionary [64] содержится 616 000 словарных статей).

Если какие-либо авторы собираются составить толковый словарь объемом, скажем, в 100-200 тыс. лексем, то они интуитивно отбирают, как выясняется, самые частотные слова. Но, как гласит семантический закон Ципфа-Гиро, высокочастотные слова как раз и являются самыми многозначными. Но Р.А. Будагову [15], по-видимому, это было неизвестно, и «по простоте душевной» он «открыл Америку наоборот», перепутав причину со следствием (что является философской и научной ошибкой!), и он, так сказать, «сел в лингвистическую лужу»!

В 1961 году вышла из печати книга «О точных методах исследования языка» [6], которая «первой ласточкой» в СССР в этой области.

Авторы сделали обзор ряда работ, в которых делались попытки применить известные формулы математики к описанию лингвистических явлений.

В частности, продемонстрировано применение некоторых простых приемов математической статистики для описания лингвистических элементов (знаков).

Например, нахождение объёма выборки N:

N = ttq / ddp

p – доля интересующих нас элементов, q – доля остальных элементов, t – константа (стремящаяся к 2 при увеличении N), d (дельта) – относительная ошибка, r* (ро) – уровень значимости (достоверности) [Заметим, что по «техническим» причинам здесь приняты обозначения: tt – это t в квадрате, dd - это d в квадрате].

Интересным является факт приведения формулы для нахождения количества разных слов L в выборке текста из N слов, предложенное П. Гиро [50]:

L = k N exp 1/2 (k = 22)

Как известно, в методике обучения иностранному языку важно знать, сколько разных слов L содержится в конкретном учебном тексте N, что позволяет осуществлять научно обоснованное дозирование материала для учащихся.

Авторы отмечают важность использования количественных закономерностей, обнаруженных при анализе частотных словарей.

Частотные словари составляются следующим образом. Отбираются тексты общим объёмом, например, в 1 млн. словоупотреблений (как это делалось при составлении больших Частотных словарей английского [56] и русского [47] языков) и подсчитывается, сколько раз встретилось каждое слово в этой выборке. Разделив эти числа на 1 млн. (то есть объём выборки), получали частотность (долю) каждого слова в тексте. Затем распределяют слова в порядке убывания их частотности, начиная с самого частотного.

Практически это означает, что если человек знает 1100 самых частотных слов иностранного языка, то он поймёт 80% слов текста (а это немело!).

Авторы этой книги [6] показали, как немецкий математик Фукс [44] получил формулу для подсчета количества слов, состоящих из одного, двух, трех и более слогов в английском и латинском языках. Вначале было подсчитано количество разносложных слов в произведениях Шекспира.

. Оказалось, что односложные слова составляют около 78%, двухсложные – 16%, трёхсложные – 4%, четырёхсложные – 1%, четырёх- и пятисложные – оставшийся 1%. У современного писателя Хаксли односложных слов оказалось 69%, двухсложных – 19%, трёхсложных – 9%, четырёхсложные – 2%, пятисложные – 1%. На графике видно, что картина распределения слов по количеству слогов почти совпадает.

Подсчёт количества слогов в словах в произведениях Цезаря и Саллюстрия (естественно, на латинском языке) показал, что их распределения очень близки: . Оказалось, что у Цезаря односложные слова составляют около 24%, двухсложные – 28%, трёхсложные –25%, четырёхсложные – 16%, пятисложные – 5% и шестисложные – оставшийся 2%. У Саллюстрия односложных слов оказалось 20%, двухсложных – 32%, трёхсложных – 30%, четырёхсложных – 13%, пятисложные – 4%, шестисложные -1%. На графике видно, что картина распределения слов по количеству слогов в этих латинских текстах сильно отличается от английских текстов, но очень сходна между двумя авторами (практически совпадает).

Всё это натолкнуло Фукса [44] сконструировать сложную эмпирическую формулу, позволяющую подсчитывать долю слов с любым количеством слогов в разных языках (индоевропейской семьи!).

В частности, авторы данной книги [6] подсчитали, что среднее количество слогов в словах русского языка равно 2,228. Расчеты по формуле Фукса [44] показывают, что, например, трёхсложных слов в русском языке 3,58%, что хорошо согласуется с результатами подсчетов по тексту.

Авторы книги [6] показывают, что ещё в 1929 году В.Томашевский [37] «количественно» анализировал ритмическую структуру стихов А. Пушкина и установил следующее: «Между числом стоп (четных слогов) в ямбе (x) и средним количеством пиррихиев в одной стихотворной строке (y) существует вполне определённое строгое соотношение $y = 0{,}28\ (x - 1)$» [6]. Полученные с помощью этой формулы величины хорошо согласуются с экспериментальными данными [6]. Далее в этой книге описывается способ нахождения степени информативности букв в тексте, применяя «метод угадывания», по формуле Шеннона [58]

$$H^* = 1 - \mathrm{Sum}\ P_i \log P_i / \log n,$$

где H^* - относительный коэффициент неоднородности распределения долей элементов по классам, P_i – доля элементов i-го класса, $\log P_i$ – двоичный логарифм доли P_i, причём номера классов «пробегают значения» i=1,2, 3,…n; Sum – знак суммирования произведений $P_i \log P_i$. Если К=1, то неоднородность распределения элементов по классам максимальна (все элементы в одном классе), если К=0, то неоднородность минимальна (то есть налицо однородность) (элементы поровну распределяются по классам).

В 1977 г. вышла книга «Математическая лингвистика» профессора Пиотровским Р.Г. с соавторами [30]. Один из соавторов – Бектаев К.Б. – был кандидатом физико-математических наук, а затем стал доктором филологических наук в 1975 г. Он-то и обеспечивал «математический» аспект этой книги.

Поэтому материал книги оказался практически недоступным для традиционных лингвистов (то есть не имеющих высшего математического образования).

Однако авторам удалось с помощью «математизированных» лингвистических рассуждений получить, например, формулу объёма словаря L(t) какого-нибудь словаря к концу периода T в виде

$$L(t) = L\ (0)\ [\exp\ (k)],$$

где L(0) – объём словаря в некий начальный момент t=0, k – коэффициент роста словаря.

Англичанин Дж. Хердан [44] предлагает оценивать долю романских заимствований в английском языке с помощью следующей формулы:

$$P = C \log N$$

Здесь р – доля романских заимствований N – объём словника, C – константа, равная 0,1 для английского языка, lg – десятичный логарифм N.

Эта формула эмпирическая, но никаких рассуждений автор не предлагает.

Попробуем оценить долю романских заимствований в словаре объёмом в 100 000 лексем. Подставив эту величину в формулу, получим р=0,5. То есть доля романской лексики в этом гипотетическом составляет половину словаря, а именно: 50 000 лексем.

В крупнейших толковых англо-английских словарях: The Oxford English dictionary [64] объёмом в 616 тыс. словарных статей и Webster's New International dictionary of the English language, содержащем 600 тыс. словарных статей, количество романских заимствований должно равняться 346 тысячам лексем (58%)!

Таким образом, знакомство лингвистов с философскими источниками общего порядка [41; 42] привели к глубокому пониманию важности философского подхода при изучении языка и речи, а философская категория «количества» начала широко использоваться в лингвистике к середине XX века [1; 27; 21; 22; 46].

В 50-х годах XX века появился ряд работ, в которых лингвисты начинают использовать более широкий спектр математических средств [2; 4; 6; 7-14; 17-19; 23; 25; 26; 29; 30; 35; 38; 43].

Источниками «математического вдохновения» служит как литература на русском языке [4; 6; 17; 18; 19; 20; 25; 26; 30; 38], так и иностранная (благо лингвисты «владеют иностранными языками») [40; 43; 44; 49-53; 58-62; 68].

Библиографический список

1. Александров П. Звуковая характеристика немецкого языка по статистическим данным // Учёные записки Казанского университета. Казань, **1911.**
2. Алексеев П.М. О семантических частотных словарях // Частотный англо-русский физический словарь-минимум / Сост. П.М. Алексеев, М.Е. Каширина, Е.М. Тарасова. М.: Воениздат, 1980. С. 265-288 с.
3. Алексеев П.М. Частотный англо-русский словарь-минимум по электронике. М.: Воениздат, 1971.
4. Андреев Н.Д. Статистико-комбинаторные методы в теоретическом и прикладном языкознании. Автореф. дис. …д-ра филол. наук. Л., 1867.
5. Андреев С.Н., Кузьмин Л.А., Луговской В.П., Сильницкий В.П. Исследование связи деривационных и синтаксических характеристик английских глаголов методом корреляционного анализа языка) // Исследование деривационной подсистемы количественным методом. Владивосток: ДВНЦ АН СССР, 1983. С.108-124..

6. Ахманова О.С., Мельчук И.А., Падучева Е.В., Фрумкина Р.М. О точных методах исследования языка. М.: Изд-во Московского ун-та, 1961. 162 с
7. Бартков Б.И. Корреляционный анализ в дериватологии // Дериватология и дериватография литературной нормы и научного стиля. Владивосток: ДВНЦ АН СССР, 1984. С. 3-27.
8. Бартков Б.И. Квантитативные методы исследования словообразовательной подсистемы современного английского языка // Аффиксоиды, полуаффиксы и аффиксы в научном стиле и литературной норме. Владивосток: ДВНЦ АН СССР, 1980. С. 117-142.
9. Бартков Б.И. Количественные методы в дериватологии (на материале немецкого языка) // Исследование деривационной подсистемы количественным методом. Владивосток: ДВНЦ АН СССР, 1983. С.3-40.
10. Бартков Б.И. Количественная семантика английских деривационных типов // Форма и содержание единиц языка и речи. Владивосток: Дальнаука, 1998. С. 3-12.
11. Бартков Б.И. Лингвистика и философия. Часть 1. Дефиниции // Дневник науки, № 16. 8 августа 2022 г. 22с.
12. Бартков Б.И. О коэффициентах сходства членов синонимических рядов (Двухмерный и многомерный простой и «взвешенный» случаи) // Структурная и математическая лингвистика. Выпуск 9. Киев: Издат-е объединение «Вища школа», 1981. С. 6-13.
13. Бартков Б.И. Очерки по количественной глоттологии и глоттографии. Казань: Изд-во «Бук», 2017. - 244 с.
14. Бартков Б.И. Системные свойства языка (научный стиль и норма) // Системный анализ научного текста. Владивосток: ДВНЦ АН СССР, 1984. С. 3-26.
15. Будагов Р.А. Что такое развитие и совершенствование языка? М.: Наука, 1977. «:; с.

16. Булыко А.Н. Большой словарь иноязычных слов. 35 тысяч слов. М.: «Мартин», 2004. ?04 с.
17. Головин Б.Н. Язык и статистика. М.: Просвещение, 1970. 191 с.
18. Денисова А.В., Миркин Б.М. Об альтернативных показателях связи, используемых при анализе биологических явлений // Биол. Науки, 1972, № 3. С. 121-128.
19. Ермоленко Г.В. Лингвистическая статистика. Краткий очерк и библиографический указатель. Алма-Ата: Казах. Гос. ун-т им. С.М. Кирова, 1970. 154 с.
20. Лакин Г.Ф. Биометрия. М.: Выс. Школа, 1980. 293 с.
21. Марков А.А. Пример статистического исследования над текстом «Евгения Онегина», иллюстрирующий связь испытаний в цепь. Известия Академии наук, серия VI, № 3, 1913.
22. Морозов Н.А. Лингвистические спектры. Санкт-Петербург: АН. Известия отд. Русского языка и словесности. АН, XX т., кн. 1-4. 1915.
23. Москович В.А. Статистика и семантика. М., 1969.
24. Новый большой русско-английский словарь (в трёх томах 300 000 слов) / Под рук. П.Н. Манкурова, М.С. Мюллера, В.Ю. Петрова. М.: «Лингвистика», 1997. Т. 1-3.
25. Носенко И.А. Начала статистики для лингвистов. М.: Выс. Школа, 1981.
26. Парчевская Д.С. Статистика для радиоэкологов. Практическое руководство по статистике и планированию экспериментов в радиоэкологии. Киев: Изд-во «Наукова думка», 1969. 115 с..
27. Петров В. Звуковая характеристика французского языка по статистическим данным // Учёные записки Казанского университета. Казань, 1911.
28. Пешковский А.М. Десять тысяч звуков // Методика родного языка, лингвистика, стилистика, поэтика. Л.: Госиздат, 1925.

29. Пиотровский Р.Г. Информационные измерения языка. Л. отд.: Наука, 1968. 116 с.
30. Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А. Математическая лингвистика. М.: Выс. Школа, 1977. 383 с.
31. Сильницкий Г.Г Регрессивный и корреляционный анализ суффиксальной сочетаемости английского глагола // Дериватология и дериватография литературной нормы и научного стиля. Владивосток: ДВНЦ АН СССР, 1984. С. 106-121..
32. Сильницкий Г.Г., Андреев С.Н., Кузьмин Л.А., Луговской В.П. О некоторых количественных методах определения мер связи между языковыми уровнями // Исследование деривационной подсистемы количественным методом. Владивосток: ДВНЦ АН СССР, 1983. С. 40-62.
33. Сильницкий Г.Г., Кристаллинский Р.Е., Андреев С.Н., Кузьмин Л.А. // О некоторых математических методах классификации лексических единиц и их признаков // Проблемы словообразования в английском и немецком языках, Смоленск: Смоленск. Гос. пед. ин-т, 1982. С. 5-19.
34. Соссюр Ф. де. Курс общей лингвистики. М.: ИЛ, 1933. 272 с.
35. Тихонов А.Н. Система русского словообразования в свете количественных данных // Исследование деривационной подсистемы количественным методом. Владивосток: ДВНЦ АН СССР, 1983. С. 61-73.
36. Тихонов А.Н. Словообразовательный словарь русского языка. В 2-х томах. – М.: Русский язык, 1985. –Т. 1. – 855 с.; Т. 2. – 887 с.
37. Томашевский Б.В. О стихе. Статьи.Л.: Прибой, 1929.
38. Тулдава Ю. Проблемы и методы квантитативно-системного исследования лексики. 1987. 204 с.
39. Улуханов И.С. О структуре статьи «Словаря словообразовательных аффиксов современного русского языкап» // Морфемология и морфемография. Владивосток: ДВО РАН, 1993. С. 88-101.

40. Фёрстер Э., Рёнц Б. Методы корреляционного и регрессионного анализа. М.: Финансы и статистика, 1983. 302 с.
41. Философия. Основные идеи и принципы: Попул. Очерк / А.И. Ракитов, В.М. Богуславский, В.Е. Чертихин, Г.И. Эзрин. М.: Политиздат,1985.368с
42. Философский энциклопедический словарь / Гл. редакция: Л.Ф. Ильичев, П.Н. Федосеев, С.М. Ковалев, В.Г. Панов - М.: Советская Энциклопедия, 1983. – 840 с.
43. Фукс В. Математическая теория словообразования // Теория передачи сообщений. М.: Изд-во иностр. Лит-ры, 1957. С. 221-247.
44. Хердан (Herdan G.) Type-Token Mathematics. S-Gravenage, 1960. 281 p.
45. Частотный англо-русский физический словарь-минимум / Сост. П.М. Алексеев, М.Е. Каширина, Е.М. Тарасова. М.: Воениздат, 1980. – 288 с.
46. Чистяков В.Ф., Крамаренко Б.К. Опыт приложения статистического метода к языкознанию. Краснодар, 1929, 69 с.
47. Частотный словарь русского языка. Около 40 000 слов. Сост. В.А.Аграев, В.В.Бородин, Л.Н.Засорина, В.М.Муратова, Э.В.Тисенко. М.: Русский язык, 1977. 836 с.
48. Eaton H.S. An English-French-German-Spanish Word Frequency Dictionary. New York, 1961.
49. Czekanowski J. Zarys metod statystycznych w zastosowaniu do antropologii // Prace Towarzystwa Naukowego Warszawskeiego. III. Wydzial nauk matematycznych i przirodniczych. № 5, Warszawa, 1913.
50. Guiraud P. Le characters statistiques du vocabulair francais. Paris, 1954. 111 p
51. Haman U. Merkmalsbestand und Verwandschaftsbeziehungen den Farinosae. Ein Beitrag zum System der Monokotyledonen. Willdenowia, 1961, Bd. 2, S. 639-768.
52. Jaccard P. The distribution of the flora in the alpine zone // New Phytol., 1912, v.11.

53. Jule G.U. The Statistical Study of Literary Vocabulary, Cambridge, 1944
54. Josselson H.H. The Russian Word Count. Detroit, 1853/
55. Kaeding F. Haeufigkeitswoerterbuch der Deutschen Sprache. Steigliz bei Berlin, 1898.
56. Kucera H., Francis W.N. Computational analysis of present-day American English. Providence, Rhode Island: Brown Univ. Press, 1967.- 491 p.
57. Merkmalsbestand und Verwandschaftsbeziehungen den Farinosae. Eib Beitrag zum System der Monokotyledonen. Willdenowia, 1961, Bd. 2, S. 639-768/
58. Shannon C.E., Weaver W. The Mathematical Theory of Communication. Univ. Illinois Press, Urbana, 1949. Pp. 117.
59. Simpson T.H. The measurement of diversity // Nature, 1941, v. 163.
60. Spearman C. Demonstration of formulae for true measurement of correlation // Amer. J. Psychol., 1907, v. 18. P. 161-167.
61. Studies in language behavior // Psychological monographs, vol. 56, No. 2 Washington, 1944.
62. Swadesh M. Lexico-statistic Dating of Prehistoric Languges.
63. The Concise Oxford Dictionary of Current English.6-th ed. Oxford,1978.763p
64. The Oxford English Dictionary. 2nd Ed. Oxford Univ. press. 1989. Vols.1-20.
65. Thorndike E., Lorge J. The Teacher's Word Book of 30,000 Words. N.Y., 1944.
66. Vander Beke G.E. French word book. N.Y.: The Macmillan Company, 1930. 188 p.
67. West M. A General Service List of English Words. London, 1953.
68. Zipf G.H. The Meaning-Frequency relationship of Words. The Journal of General Psychology, v. 33, 1945. 364 p.

The End

Оригинальность 98%