ISSN 0494-7304 0207-4435

TARTU ÜLIKOOLI TOIMETISED

УЧЕНЫЕ ЗАПИСКИ ТАРТУСКОГО УНИВЕРСИТЕТА

ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS

872

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

1989

QUANTITATIVE LINGUISTICS AND AUTOMATIC TEXT ANALYSIS

TARTU 1989

TARTU ÜLIKOOLI TOIMETISED
УЧЕНЫЕ ЗАПИСКИ ТАРТУСКОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS
Alustatud 1893.a. VIHIK 872 ВЫПУСК Основаны в 1893.г.

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

1989

QUANTITATIVE LINGUISTICS AND AUTOMATIC TEXT ANALYSIS

TARTU 1989

Kogumik "Kvantitatiivlingvistika ja tekstide automaatanalüüs" ilmub Tartu Ülikooli rakendus- ja arvutuslingvistika uurimisgrupi iga-aastase väljaandena alates 1985.a. (jätkates sarja "Töid keelestatistika alalt" I - X, mis ilmus 1976-1984). Käesolevas viiendas (15-ndas) väljaandes (1989) on avaldatud kõrgkoolide vahelise probleemrühma "Tekst interdistsiplinaarse uurimise objektina" liikmete artiklid.

Сборник "Кватитативная лингвистика и автоматический анализ текстов" публикуется Группой прикладной и компьютерной лингвистики Тартуского университета начиная с 1985 г. (сборник является продолжением серии "Труды по лингвостатистике" I - X, 1976-1984 гг.). Настоящий 5-й (15-й) выпуск (1989 г.) содержит статьи членов Межвузовской проблемной группы "Текст как объект междисциплинарных исследований".

The collections "Quantitative Linguistics and Automatic Text Analysis" appears annually since 1985, edited by the members of the Applied and Computational Linguistics Group at Tartu University (Estonia). The collections continue the series "Papers on Linguo-Statistics" (1976-1984).
The present issue No. 5 (15) contains investigations by members of the All-Union Research Group "Text as an object of interdisciplinary investigations".

МНОГОМЕРНЫЙ АНАЛИЗ РАЗНОУРОВНЕВЫХ ПРИЗНАКОВ АНГЛИЙСКИХ АФФИКСАЛЬНЫХ ГЛАГОЛОВ

С.Н. Андреев, Ю.А. Тулдава

В число важнейших задач лингвистики входит проблема исследования структурной организации системы языка, выявление связей и зависимостей между языковыми элементами и признаками. Оптимальной стратегией для решения этой задачи, по мнению многих исследователей, является построение классификации, или группировки, данных (Панова Н.С., Шрейдер Ю.А., 1975; Миркин Б.Г., 1985, с. 161). Такой подход рассматривается как самый простой и естественный способ анализа информации, отражающий логику эмпирического исследования.

Целью данной работы является группировка ряда фонетических, деривационных, семантических и синтаксических признаков английских производных аффиксальных глаголов для выявления структурной организации деривационной подсистемы глаголов современного английского языка.

Материал исследования составили 885 производных глаголов, которые включают в свой состав 12 наиболее частотных деривационных аффиксов: 8 префиксов (be-, de-, dis-, en-, out-, over-, re-, un-) и 4 суффикса (-ate, -fy, -ize, -en). Этот список был получен в результате сплошной выборки из словаря Hornby A.S. Oxford Advanced Learner's Dictionary of Current English. 1980, 3rd ed.

Каждый глагол данного списка характеризуется (описывается) тридцатью одним альтернативным (дихотомическим) признаком, список которых приводится ниже. Иными словами, относительно каждого глагола устанавливается наличие или отсутствие каждого из 31 признака. В скобках приводятся примеры, после двоеточия - сокращенное название. (1) Начальная фонема - согласная: "СГЛ (see, beg, gaze). (2) Конечная фонема - согласная: СГЛ" (read, own). (3) Наличие двух слогов: 2СЛ (insult, object). (4) Наличие трех слогов: 3СЛ (reconvene, specify). (5) Наличие четырех слогов: 4СЛ (remonetize, reciprocate). (6) Ударность первого слога: 1УД (have, punish). (7) Ударность второго слога: 2УД (efface, inscribe). (8) Ударность третьего слога: 3УД (reconvene). (9) Переходность: ТРНЗ (John built a house). (10) Сочетаемость с косвенным дополнением: КСВ (John gave Ann a book). (11) Сочетаемость со вторичным предикатом: 2СКЗ (John saw his friend put the key in

the lock and open the door). (I2) Сочетаемость с придаточным предложением: ПРИД (John said that he would go to London). (I3) Непереходность: ИНТР (John is sleeping). (I4) Глагол относится к древнеанглийскому периоду: ДРЕВ (bake, like). (I5) Глагол относится к среднеанглийскому периоду (catch, quit). (I6) Глагол относится к новоанглийскому периоду: НОВ (ameliorate). (I7) Глагол является производящим: ДЕР. (I8) Глагол сочетается с префиксом: ПФ (enchant - disenchant). (I9) Глагол сочетается с суффиксом: СФ (decide - decision). (20) От глагола образуется существительное: СУЩ (govern - government). (2I) От глагола образуется глагол: ГЛ (stir - bestir). (22) От глагола образуется прилагательное: ПЛГ (defend - defensive). (23) Глагол имеет соотносимое с ним по конверсии существительное (независимо от направления деривации в такой глагольно-именной паре): КОНВ (dance - dance N.). (24) Глагол является каузативным: КЗ. (25) Глагол является некаузативным: НКЗ. (26) Пространственные соотношения: ПРОС (go, move, slope, orient, fly). (27) Информация: ИНФР (say, talk, see, smell). (28) Положительная оценка: ОЦ+ (approve, praise). (29) Негативная оценка: ОЦ- (criticise, denounce). (30) Глагол имеет постериорный компонент значения: ПОСТ (promise). (3I) Глагол имеет антериорный компонент значения: АНТР (punish). Здесь используется семантическая система понятий, предложенная Г.Г. Сильницким (Сильницкий Г.Г., I986).

Данные признаки были отобраны в связи с тем, что они оказались наиболее релевантными в плане межуровневых соотношений и составляют ядерную часть признаковой системы английских глаголов (Сильницкий Г.Г. и др., I989).

В результате такой процедуры получена таблица "объект - признак" I2 x 3I, в которой строками являются аффиксальные типы глаголов, а столбцами - их фонетические, деривационные, синтаксические и семантические параметры. На пересечении строк и столбцов указывается число глаголов данного аффиксального класса, обладающих данным признаком.

Имеются различные методы достижения разбиений исследуемых объектов (Энслейн К., I986); центральное место среди них занимают кластерный и факторный виды анализа.

Кластер-анализ.

Одним из основных методов, направленных на группировку объектов или признаков является кластерный анализ. Этот метод обычно включает два основных этапа: преобразование исходных многомерных данных в данные о близости, преобразование этих

данных с помощью ряда процедур в данные о кластерах. Под термином "близость" понимается сходство, различие, корреляция или любая другая переменная, используемая в качестве меры сходства или расстояния между каждыми двумя объектами из всего множества объектов, подлежащих группировке. Различные виды кластерного анализа в лингвистике использовались в целом ряде работ, как правило, при анализе текстового материала (Мартыненко Г.Я., 1988; Тулдава Ю.А., 1987; Лийв Х., Тулдава Ю., 1987; Шайкевич А.Я., 1979 и др.). В данной работе кластер-анализ используется для классификации элементов языковой системы.

В качестве матрицы близости для всех 465 возможных пар признаков (31 x (31 - 1) / 2) был использован коэффициент линейной корреляции Пирсона-Бравэ

$$r = \frac{\sum (x_1-\bar{x})(y_1-\bar{y})}{(n-1)S_x S_y} \cdot$$ (Методика..., 1968).

Все вычислительные работы были проведены в ВЦ ТГУ по разработанным там программам на ЭВМ ЕС-1060. Корреляционная матрица приводится в табл. 1.

При конструировании кластер-системы на основании анализа данных этой таблицы, преобразованных по формуле 100 (r + 1) / 2, использовалась агломеративная процедура послойной кластеризации (Лийв Х., Тулдава Ю., 1987). Число кластеров, в которые может входить один и тот же элемент, не превышает двух. Результаты данной процедуры приводятся на рис. 1 а, б. На уровне h_1 = 0,8941 выделяются 16 кластеров, указанных на рис. 1а: два трехэлементных, один двухэлементный, и один одиннадцатиэлементный и двенадцать одноэлементных кластеров.

На уровне h_2 = 0,8770 происходит некоторое увеличение числа элементов выделенных кластеров и уменьшение одноэлементных кластеров (рис. 2б).

Факторный анализ.

Факторный анализ представляет собой метод, который так же как и кластерный анализ предназначен для сжатия информации, содержащейся в корреляционной матрице. Он позволяет сгруппировать параметры таким образом, чтобы их группы в наибольшей степени отражали сущность исследуемого явления и вскрывали факторы, влияющие на образование этих групп. В основе факторного анализа лежит следующая гипотеза, сформулированная Г. Харманом: "Наблюдаемые или измеряемые параметры

являются лишь косвенными характеристиками изучаемого объекта или явления, на самом же деле существуют внутренние (скрытые, не наблюдаемые непосредственно) параметры или свойства, число которых мало и которые определяют значения наблюдаемых параметров. Эти внутренние параметры принято называть факторами" (Харман Г., 1972, с. 7).

В данной работе факторный анализ проводится в рамках метода главных факторов (Харман Г., 1972, с. 151). В лингвистике данная методика была использована в работе (Тулдава Ю., 1976). Результаты вычисления приводятся в таблице 2, которая представляет собой факторную матрицу, полученную после вращения факторов по методу "варимакс" (Харман Г., 1972, с. 326). В таблице представлено пять основных факторов. Каждый из 31 признака характеризуется системой нагрузок (весов) этих факторов. Так, коэффициент 0,79 в первой строке и первом столбце означает положительную зависимость между признаком "СГЛ и фактором 1. Чем выше эта корреляция, тем в большей степени параметр "наполнен" данным фактором и тем в большей степени является мерой этого фактора.

Путем сравнения факторных нагрузок в столбцах факторной матрицы при одновременном учете характера параметра можно прийти к определенным гипотезам о природе полученных факторов. Анализ факторов обычно завершается тем, что дается соответствующее наименование. В нашем случае, однако, вряд ли целесообразно стремиться к такому исчерпывающему интерпретационному анализу в связи с тем, что к исследованию привлекаются разноуровневые признаки, информации о соотношении которых для деривационной подсистемы пока имеется чрезвычайно мало. Поэтому ограничимся на данном этапе исследования лишь выделением для каждого фактора характеризующих его элементов, факторные нагрузки которых относительно выше, чем у других параметров.

Относительно более высокие положительные факторные нагрузки были зафиксированы у следующих признаков.

Фактор 1: СГЛ", 3СЛ, 4СЛ, ТРНЗ, НОВ, ДЕР, СФ, СУЩ, КЗ, ИНФР, ОЦ-, ПОСТ, "СГЛ.

Фактор 2: 2УД, ПРИД, КОНВ, ПРОС, АНТР.

Фактор 3: 2СЛ, 2СКЗ, КОСВ.

Фактор 4: 1УД, ИНТР, ПФ, ГЛ, НКЗ.

Фактор 5: 3УД, ДРЕВ, ОЦ+.

Последний фактор является биполярным, так как он помимо положительных характеризуется также и отрицательными факторны-

ми нагрузками на параметрах КОСВ и 2УД.

Результаты факторизации удобно изобразить графически, что позволяет получить новую информацию о структуре признакового пространства данного вида (Тулдава Ю., 1976, с. 134). Каждый признак может быть геометрически интерпретирован как точка в некотором пространстве, которое имеет столько измерений, сколько было выделено факторов. Однако практически удобно строить пространственную модель для двухмерных случаев, принимая во внимание только два фактора и их проекции на оси координат. Для иллюстрации графического изображения приводим примеры с первым и вторым, а также первым и третьим факторами. На рис. 2 осями координат являются факторы I и 2, на рис. 3 - факторы I и 3. Координаты точек определяются факторными нагрузками исходных признаков. Расположение точек на рис. 2 свидетельствует о распадении совокупности исходных параметров на группы: 2СЛ (на рис. точка 3) - СРЕД (I5); ПРОС (26) - КОНВ (23) - АНТР (3I) - ПРИД (I2) - 2УД (7); ОЦ+ (28) - 3УД (8) - 2СКЗ (II). Четвертую "группу" составляет один признак ДРЕВ (I4), а в пятую входят все остальные признаки.

В пространстве I и 3 факторов признаки разбиваются на следующие группы: СРЕД (I5) - 2СЛ (3); ДРЕВ (I4); 2СКЗ (II); КОНВ (23) - АНТР (3I) - ПРИД (I2) - 3УД (8) - ПРОС (26) - ОЦ+ (28) - 2УД (7); НКЗ (25) - КОСВ (I0) - ИНТР (I3) - ПОСТ (30) - СГЛ" (2) - "СГЛ (I); КЗ (24) - ТРНЗ (9) - НОВ (I6) - СФ (I9) - ДЕР (I7) - ИНФР (27) - СУЩ (20); ГЛ (2I) - ПФ (I8) - IУД (6) - 4СЛ (5) - 3СЛ (4) - ПРИЛ (22) - ОЦ- (29).

Использование в качестве пространственной координаты фактора 3 вместо фактора 2 вносит некоторые изменения в группировку признаков. Так признак 2СКЗ оказывается изолированным в этом случае, а признаки ОЦ+ и 3УД, которые входили с ним в одну и ту же группу, теперь объединяются с признаками другой группы (КОНВ, АНТР и др.).

Сопоставление результатов факторного и кластерного видов анализа показывает, что первый фактор почти полностью совпадает с наибольшим по числу элементов кластером (I). Второй фактор также почти полностью совпадает с кластером (2), а четвертый фактор соотносится с кластерами (4) и (3). Факторы Ш И У объединают признаки, которые при кластерном анализе выделялись в одноэлементные кластеры. Таким образом значительные различия наблюдаются только для двух последних случаев. В целом же можно констатировать высокую степень устойчивости получаемых разбиений признаков, что свидетельст-

вует о существенности выявляемых структурных особенностей признаковой системы.

Полученные объединения признаков могут рассматриваться как база для дальнейшей классификации конкретных аффиксальных глаголов. Соотнесение глаголов с данными комплексами признаков позволит выделить лексические единицы, которые относятся к сфере "ядра" того или иного комплекса, а также глаголы, совмещающие признаки различных "архетипов". Иными словами здесь речь может идти о степени прототипичности конкретной лексики относительно полученных объективным образом "идеальных" образцов - объединений разноуровневых признаков.

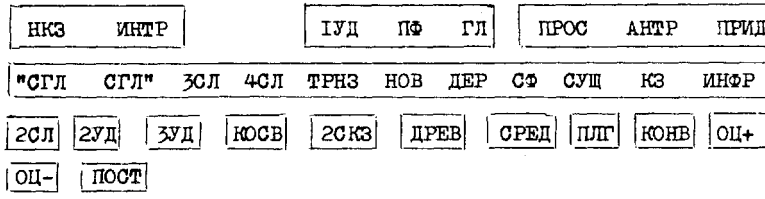

Рис. Iа. Группировка признаков методом кластерного анализа при h_1 = 0,894I

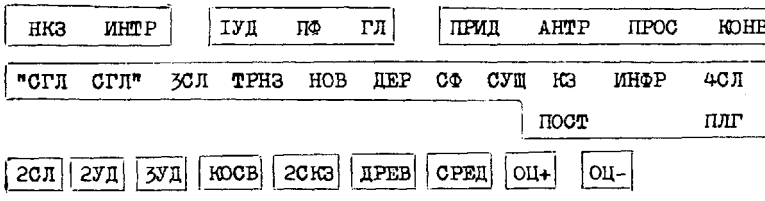

Рис. Iб. Группировка признаков методом кластерного анализа при h_2 = 0,8770

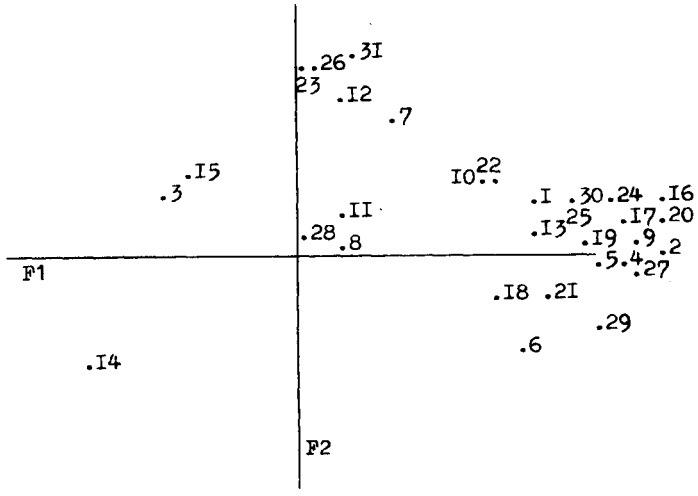

Рис. 2. Представление 31 признака в плоскости двух первых факторов

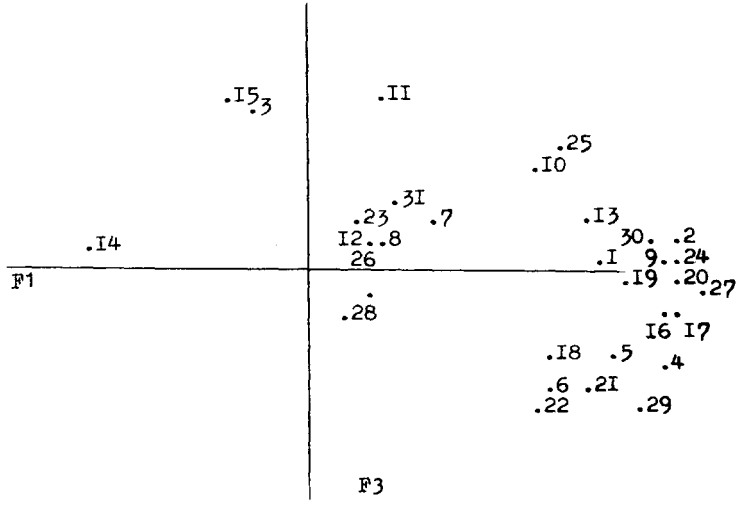

Рис. 3. Представление 31 признака в плоскости первого и третьего факторов

Факторная матрица признаков английских аффиксальных глаголов (варимакс-решение)

Признаки	Факторные нагрузки				
	F1	F2	F3	F4	F5
ТРНЗ	0,967	0,0	0,0	0,0	0,0
СУЩ	0,967	0,0	0,0	0,0	0,0
СФ	0,965	0,0	0,0	0,0	0,0
ДЕР	0,963	0,0	0,0	0,0	0,0
НОВ	0,956	0,0	0,0	0,0	0,0
КЗ	0,937	0,0	0,0	0,0	0,0
ИНФР	0,934	0,0	0,0	0,0	0,0
СГЛ"	0,923	0,0	0,0	0,0	0,0
ЗСЛ	0,910	0,0	0,0	0,0	0,0
4СЛ	0,859	0,0	-0,384	0,0	0,0
ОЦ-	0,798	-0,316	0,0	0,0	0,0
"СГЛ	0,791	0,289	0,0	0,0	0,0
ПОСТ	0,768	0,286	0,0	0,0	-0,321
ГЛ	0,670	0,0	0,0	0,597	-0,274
ІУД	0,656	-0,326	0,0	0,621	0,0
ИНТР	0,645	0,0	0,0	0,633	0,0
ДРЕВ	-0,541	-0,410	0,0	0,0	0,433
ЮСВ	0,538	0,336	0,468	0,0	-0,426
ПЛГ	0,537	0,368	-0,394	0,415	0,0
ПРОС	0,0	0,939	0,0	0,0	0,256
АНТР	0,0	0,879	0,328	0,0	0,0
ПРИД	0,0	0,875	0,0	0,0	-0,268
ЮНВ	0,0	0,750	0,458	0,0	0,300
2УД	0,362	0,633	0,0	-0,393	-0,462
2СЛ	-0,330	0,346	0,750	0,0	0,0
2СКЗ	0,0	0,256	0,726	0,0	-0,315
СРЕД	-0,303	0,449	0,652	0,0	0,0
НКЗ	0,604	0,0	0,379	0,649	0,0
ПФ	0,622	0,0	0,0	0,628	-0,253
ЗУД	0,0	0,0	0,0	-0,314	0,830
ОЦ+	0,0	0,0	0,0	0,0	0,828
	14,052	4,862	2,977	2,929	2,704

Л И Т Е Р А Т У Р А

Лийв Х., Тулдава Ю. О классификации текстов с помощью кластер-анализа // Учен. зап. ТГУ. Тарту, 1987. Вып. 777. - С. 55-68.

Мартыненко Г.Я. Основы стилеметрии. - Л.: Изд-во Ленинградского ун-та, 1988. - 74 с.

Методика и техника статистической обработки первичной социологической информации. - М.: Наука, 1968. - 326 с.

Миркин Б.Г. О математическом аппарате метода группировок в современных социально-экономических исследованиях // Анализ нечисловой информации в социально-экономических исследованиях. М.: Наука, 1985. - С. 161-169.

Панова Н.С., Шрейдер Ю.А. Принцип двойственности в теории классификации // Научно-техническая информация. ВИНИТИ, Серия 2. 1975. № 10. - С. 3-10.

Сильницкий Г.Г. Семантические классы глаголов в английском языке. Смоленск, 1986. - 112 с.

Сильницкий Г.Г., Андреев С.Н., Кузьмин Л.А., Кусков М.И. Соотношение глагольных признаков различных уровней в английском языке. - Минск: Наука и техника. 1989.

Тулдава Ю. Опыт квантитативного анализа художественного стиля // Учен. зап. ТГУ. Вып. 396. Тарту, 1976. - С. 122-141.

Тулдава Ю. Проблемы и методы квантитативно-системного исследования лексики. - Таллинн: Валгус, 1987. - 204 с.

Харман Г. Современный факторный анализ. М., 1972. - 486 с.

Шайкевич А.Я. Дифференциация статистических классификации текстов // Исследования по общему и прикладному языкознанию. Тарту, 1979. - С. 100-106.

Энслейн К. Введение в область статистических вычислений // Статистические методы для ЭВМ. М.: Наука, 1986. - С. 11-26.

MULTILEVEL ANALYSIS OF ENGLISH AFFIXAL VERBS

Sergei Andreev, Juhan Tuldava

S u m m a r y

This article deals with the measuring of mutual connection between 31 multilevel (phonetic, morphologic, syntactic, semantic) features of English affixal verbs. On the basis of a correlation (contingency) matrix cluster-analysis and factor-analysis of the features have been performed.

СТРУКТУРИЗАЦИЯ ХУДОЖЕСТВЕННОЙ ПРОЗЫ С ИСПОЛЬЗОВАНИЕМ ЭВМ
2. ДЕТАЛИЗАЦИЯ СТРУКТУРИРОВАННОГО ТЕКСТА

О.Н.Гринбаум

В данной статье продолжено изложение формально-пунктуационного метода структуризации художественной прозы (Гринбаум О.Н.,1988). Рассмотрены метод и алгоритмы маркировки абзацев и предложений по типам письменной речи, детализации вилов прямой речи, описана система диалоговой корректировки структурированного текста. Представлены основные результаты детализации текста романа М.Ю.Лермонтова 'Герой нашего времени'.

Маркировка абзацев и предложений по типам письменной речи. Филолого-статистическому исследованию художественной прозы должно, на наш взгляд, предшествовать разграничение речевого материала на две категории - авторскую речь и чужую речь. Чужая речь включает прямую речь, диалоги, прямую речь внутри прямой речи и другие виды неавторского повествования. Имея в своем распоряжении только лишь формальные средства - знаки пунктуации и символы графической системы языка, мы предприняли попытку найти такие изобразительные средства,которые позволили бы автоматически (и одновременно с процессом структуризации) вести разграничение предложений по типам письменной речи. Представленный ниже метод автоматической маркировки структурных элементов художественной прозы - абзацев и предложений - базируется на строгом описании средств оформления прямой речи и ее пространственном расположении на страницах печатных изданий.

В алгоритме маркировки основным структурным элементом текста считается абзац : все предложения абзаца, последовательно вычленяемые из текста программой структуризации, подвергаются анализу на предмет наличия в них прямой речи; решение отнести тот или иной абзац к одному из типов письменной речи принимается по завершении анализа всех предложений данного абзаца. Сделано это с одной целью - разработать программу, обеспечивающую однородность совокупности лингвистических данных, подвергаемых статистической обработке, поскольку любые виды прямой речи, включенные в состав авторских абзацев, должны изменять маркер этих абзацев и исключать тем самым их участие в дальнейших исследованиях авторской речи.

Такие абзацы могут впоследствии составить исходный материал для углубленного анализа стилистических приемов и особенностей преломления авторской манеры письма в речи персонажей (Кожевникова Н.А., 1977).

Для маркировки абзацев и предложений будем пользоваться следующей системой обозначений :

%А% - авторский абзац(точнее - его первое предложение);

%П% - абзац прямой речи (его первое предложение);

%а% - предложение авторского абзаца (не первое);

%п% - предложение абзаца прямой речи (не первое).

Тогда появление элементов прямой речи в любом предложении авторского абзаца должно изменять тип этого абзаца (маркер первого предложения) с %А% на %П%, а каждое из последующих предложений, входящих в состав данного абзаца,должно метиться как %п%.

Пример 1. Рассмотрим последовательность маркировки следующего абзаца :

> В этот вечер Казбич был угрюмее, чем когда-нибудь, а я заметил, что у него под бешметом надета кольчуга. "Недаром на нем эта кольчуга, -подумал я, ух он, верно,что-нибудь замышляет".
>
> (Лермонтов М.Ю. Герой нашего времени)

В первом предложении этого абзаца нет элементов прямой речи, следовательно, оно после вычленения из текста будет помечено маркером %А%. Второе предложение - предложение с прямой речью, поэтому его маркер получит значение %п%, а маркер первого предложения будет изменен на %П%.

Перейдем теперь к описанию средств графического оформления и пространственного расположения элементов прямой речи.

Как известно (см.,напр.,: Шапиро А.Б.,1974), существуют два варианта построения текстов с прямой речью :

а) прямая речь оформляется в виде отдельных реплик участников разговора; в тексте такие реплики начинаются с новой строки;

б) прямая речь входит в состав абзаца авторского повествования; в этих случаях прямая речь обычно заключается в кавычки.

Построение алгоритма маркировки абзацев для первого варианта не вызывает особых затруднений : наличие абзацного отступа, за которым всегда следует тире, однозначно определяет всю ситуацию в целом. Первое предложение такого абзаца,

как мы уже отмечали, метится маркером %П%, а все последующие – маркером %п%.

Здесь необходимо сделать небольшое уточнение, и мы об этом уже говорили : ситуация < АО + тире >, зафиксированная в начале строки, требует дополнительной проверки на отсутствие двоеточия в конце предыдущей строки; поэтому при маркировке абзацев сначала проверяется наличие этих двух условий, а лишь затем первое предложение метится как %П%.

Уточним : ситуация < АО + тире > является необходимым и достаточным условием фиксации в тексте прямой речи, но потребность в детализации прямой речи по ее типам (см. ниже) приводит к необходимости детализировать и сам алгоритм маркировки. Поэтому, если в конце строки стоит двоеточие, а следующая строка начинается с < АО + тире >, то это является свидетельством появления в тексте 'смешанной' речи,т.е. прямой речи внутри абзаца авторской речи и метится как %С%.

Второй вариант построения текстов с прямой речью использует кавычки – ключевой элемент для фиксации и выделения прямой речи. И если бы те же кавычки не использовались для обозначений в тексте заимствований (названий, цитат, пословиц, поговорок и др.), то алгоритм маркировки был бы для этого варианта просто тривиален.

Пример 2.

> Петушков встрепенулся, – а солдат вытянулся,пожелал ему "здравья" и вручил ему большой конверт, запечатанный казенной печатью.
>
> (Тургенев И.С. Петушков)

Это предложение, конечно же, не должно помечаться как %П%, ибо в нем нет прямой речи; алгоритм маркировки должен отличать такие случаи и относить их к авторской речи.

Для того, чтобы реализовать это требование, можно было с самого начала двигаться в двух направлениях : либо проанализировать и выявить характерные (формальные, отличительные) признаки для предложений с заимствованиями, либо то же самое уточнить и формализовать для предложений с прямой речью.Первый способ, кажущийся вначале более предпочтительным, на самом деле алгоритмизации не поддается. Причина здесь одна – отсутствие каких-либо ограничений на использование в тексте цитат и прочих заимствований, что в отношении поиска отличительных пунктуационных особенностей делает задачу бесперспективной.

Поэтому выбран был второй путь – проанализированы и

формализованы правила графического оформления прямой речи с использованием кавычек. Для того, чтобы описать эти правила в сжатом виде, обратимся к уже использованным ранее формализмам : пусть N -ситуация описывает последовательность символов, расположенных в начале предложения, F -ситуация описывает последовательность символов, находящихся в конце предложения, а S -ситуация (это новое обозначение) - наличие определенных символов в середине предложения. Знаком Z по-прежнему будем обозначать любой из разграничительных знаков препинания $Z=('.','?','!')$, символом E - любую прописную букву, знаком 'V' - логическую операцию 'ИЛИ',а аббревиатурой 'nil' - отсутствие кавычек в N- или F-ситуациях.Тогда правила фиксации прямой речи, оформленной с помощью кавычек, могут быть сведены к четырем случаям, представленным в Табл.2.

Таблица 2

Формализованное представление правил выявления прямой речи

F - ситуация (конец предлож.)	$<Z"> \vee <"Z>$			nil
N - ситуация (начало предл.)	$<"E>$	nil		$<"E>$
S-ситуация (середина предл)	$<-...->$	$<:"E>$	$<",->$	$<Z"->$
ВАРИАНТ	1	2	3	4

Вариант 1 реализует тот случай, когда прямая речь разрывается словами автора на две части. Здесь S -ситуация должна содержать два тире, разделенных некоторым текстом.

Пример 3 (вариант 1, $F = <Z">$).

"Может быть, - подумал я, - ты оттого именно меня
и любила:радости забываются, а печали никогда..."

(Лермонтов М.Ю. Герой нашего времени)

Следующий, второй вариант реализует тот случай, когда прямая речь находится в конце предложения и следует за словами автора.

Пример 4 (вариант 2, $F = < "Z >$).

Староста сперва проворно соскочил с лошади, поклонился барину в пояс, промолвил : "Здравствуйте, батюшка Аркадий Павлович".

(Тургенев И.С. Бурмистр)

Варианты 3 и 4 реализуют случаи, когда прямая речь находится в начале предложения, а слова автора следуют за ней.

Пример 5 (вариант 4).

"Скорей, скорей в город за лекарем!" - кричал Владимир.

(Пушкин А.С. Дубровский)

Те же правила выявления прямой речи применимы и в более сложных случаях, а именно тогда, когда прямая речь состоит не из одного, а из нескольких предложений. При определении типа абзаца те предложения, которые расположены внутри кавычек, рассматриваются как одно целое, а знаки препинания '.', '?' и '!' в расчет не принимаются.

Пример 6 (вариант 1).

"Может быть, мы никогда больше не увидимся, - сказал он мне. - Перед разлукой я хотел бы с вами об'ясниться".

(Пушкин А.С. Выстрел)

Здесь прямой речью об'единены два предложения. Анализ текста, расположенного между кавычками, приводит к следующему результату : $f = <$ кавычки + '.' $>$, $\pi = <$ кавычки + 'М' $>$, $S = <$ тире + 'сказал он мне' + тире $>$; точка после слова 'мне' игнорируется. Поэтому факт наличия прямой речи будет зафиксирован согласно правила 1 (см. табл.2).

Пример 7 (вариант 3).

"Наплюй на него, бабонька. Была бы шея, а ярмо будет", - с нескрываемым сожалением посоветовала одна.

(Шолохов М.А. Тихий Дон)

В этом примере $f = nil$, $\pi = <$ кавычки + 'Н' $>$, а $S = <$ кавычки + ',' + тире $>$, что соответствует варианту 3 таблицы 2. Размеченный текст для данного примера будет состоять из двух предложений :

1. %П% "Наплюй на него, бабонька.
2. %п% Была бы шея, а ярмо будет", - с нескрываемым сожалением посоветовала одна.

Разработанные и представленные выше формальные правила выявления прямой речи охватывают три наиболее употребитель-

ных варианта построения предложений с вводящими словами автора, а именно, когда слова автора а) предшествуют прямой речи; б) следуют за прямой речью и в) включаются в прямую речь, разделяя ее на части.

Существует, как известно, еще один вариант, когда вводящие слова автора включают в себя прямую речь. Такой способ построения предложений с прямой речью используется относительно редко; тем не менее он тоже был подвергнут формализации и оказалось, что некоторых изменений в уже имеющихся правилах достаточно для обработки всех вариантов построения предложений с вводящими словами автора.

Пример 8.

> На вопрос мой : "Жив ли старый смотритель?" никто не мог дать мне удовлетворительного ответа.
>
> (Пушкин А.С. Станционный смотритель)

В начале этого предложения кавычки отсутствуют, т.е. N-ситуация не фиксируется ($n = nil$); в середине предложения распознается ситуация $S = <$: "Ж $>$, а ситуация $< Z$ " $>$, которую мы (см. вариант 2 табл.2) должны были бы искать и обнаружить в конце предложения, находится здесь в его середине.

В следующих двух примерах ситуация аналогична, но N-ситуация находится не в начале, а в середине предложения.

Пример 9.

> Я только тогда выпрямился и подумал : "Зачем это отец ходит ночью по саду?" - когда опять все утихло вокруг.
>
> (Тургенев И.С. Первая любовь)

Пример 10.

> И только когда он шептал : "Мама! Мама!" - ему становилось как будто легче...
>
> (Чехов А.П. Степь)

Этих примеров вполне достаточно, чтобы убедиться в следующем : в рассматриваемой нами конструкции вводящие слова автора, находящиеся левее прямой речи, отделяются от нее двоеточием. Следовательно, вычленение предложений с вводящими словами автора, которые включают в себя прямую речь, должно производиться при наличии следующей ситуации (вариант 5):

$$(S_1 = <\ :\ \text{"Е}\ >)\ \&\ ((S_2 = <\ \text{"},\ >)\ \vee\ (S_3 = <Z\ \text{"}\ >)),$$

где знаком '&' обозначена логическая операция 'И', а ситуация S_1 должна предшествовать появлению в тексте ситуации S_2 или ситуации S_3 .

Таким образом, мы описали еще один, пятый по счету, ва-

риант фиксации прямой речи, если последняя заключается в кавычки. Соответствующее этому варианту правило формализованного выявления прямой речи выглядит следующим образом :

$$(F = nil) \;\&\; (N = nil) \;\&\; S_1 \;\&\; (S_2 \vee S_3) \;.$$

Представим теперь результаты проверки программы маркировки абзацев и предложений на материале романа М.Ю.Лермонтова 'Герой нашего времени'.

В тексте романа выделено 447 авторских абзацев и 723 абзаца прямой речи (каждая реплика диалога – это новый абзац прямой речи). В абзацах авторской речи зафиксировано 1235 предложений, а в абзацах прямой речи – 1674 предложения. Вариант 5 (слова автора включают в себя прямую речь) встретился в тексте только один раз, ошибок при выделении прямой речи не обнаружено, предложения с заимствованиями (например "оказия", "черта","последняя туча рассеянной бури" и другие) обработанны программой маркировки правильно.

Эти результаты показывают, что разработанный нами алгоритм маркировки абзацев и предложений обладает достаточно высокой 'лингвистической' надежностью. Критически подходя к оценке качества полученных результатов, следует еще раз отметить те исторически сложившиеся факторы (нечеткость пунктуационной системы языка, авторская вольность при расстановке знаков препинания и др.), которые не позволяют гарантировать абсолютную надежность формальных методов структуризации художественной прозы. Поэтому мы вновь подтверждаем наш тезис о том, что :

1) ошибки при формально-пунктуационном членении сплошного текста на абзацы и предложения, также как и при их маркировке не могут быть полностью исключены;

2) устранение этих ошибок должно вестись исследователем-филологом в режиме диалога 'человек-ЭВМ';

3) формальная структуризация художественных текстов и маркировка абзацев призваны и в решающей степени облегчают этот ручной, поистине титанический труд;

4) облегчение труда исследователя, достигаемое за счет применения разработанного и описанного здесь комплекса программ таково, что позволяет перевести задачу использования ЭВМ в филолого-статистических исследованиях авторской и неавторской манеры письма из области желаемого в разряд реальных практических действий.

Детализация видов прямой речи. Внимательное рассмотрение вопросов, связанных с появлением прямой речи в художествен-

ных текстах,подводит нас к необходимости продолжить ее дальнейшую конкретизацию. В самом деле, первый вид прямой речи – прямая речь внутри авторского абзаца – встречается в текстах художественной прозы весьма и весьма часто. Такое сочетание авторской и прямой речи внутри одного абзаца позволяет оживить повествование, приблизить его к устной речи,а при необходимости – легко изменить направленность изложения (Кожевникова Н.А.,1977).

Другой вид прямой речи – диалог персонажей, используя который автор напрямую сталкивает своих героев, вкладывает в их слова различные оттенки и отношения к происходящим событиям, друг к другу, к историческим и культурным традициям и т.д. И если первый вид прямой речи может быть в большинстве случаев заменен косвенной речью, что, скорее всего, приведет к потере образности и эмоциональной окраски повествования,но сохранит смысловой потенциал высказываний,то в диалогической речи подобная замена практически исключена.

Итак, мы различаем и, соответственно, по-разному маркируем :

а) собственно прямую речь, т.е. прямую речь без авторских слов, т.е. речь диалогическую; для нее маркерами являются %П% и %п%;

б) 'смешанную' речь, т.е. прямую речь внутри авторских абзацев; все предложения таких абзацев метятся как %с%, а первое – как %С%;

в) 'вложенную' прямую речь, т.е. прямую речь внутри абзаца собственно прямой речи, которая маркируется как %В% и %в%.

Все три вида прямой речи характеризуют, на наш взгляд, особые стилистические приемы, которыми в той или иной степени владеет каждый автор и, более того, которые должны изучаться каждый в отдельности и лишь затем – все вместе, в общей своей совокупности как интегрированный набор методов, приемов, средств авторской манеры письма.

В связи с этим алгоритм маркировки абзацев и предложений включает в себя отдельный блок, накапливающий сведения о типе предложений одного абзаца и, при необходимости, изменяющий их маркеры по следующей схеме :

а) если в абзаце, первоначально имевшем маркер %А% , встретится прямая речь, то все маркеры предложений этого абзаца заменяются на %с%, а для первого предложения – на %С%.

б) если в абзаце, имевшем маркер %П%, встретится пря-

мая речь, то все маркеры должны измениться на %в% и %В%.

Пример 11.

> Я отдал бабе корзинку. Двери заперли, но все еще слышались рыдания и крик: "Нина!" Я сел в пролетку и приказал извозчику ехать не спеша к Невскому.
>
> (Чехов А.П. Рассказ неизвестного человека)

Маркер первого предложения, вначале имевший значение %А%, будет - после вычленения из текста второго предложения - заменен на %С%, а все последующие предложения абзаца будут помечены маркером %с%.

Скажем теперь о результатах проверки алгоритма детализации видов прямой речи : в тексте романа 'Герой нашего времени выделено 619 абзацев собственно прямой речи(1153 предложений), 85 абзацев 'смешанной' речи (422 предложения) и 19 абзацев вложенной прямой речи (99 предложений).Ошибок при работе алгоритма детализации не обнаружено; что же касается дальнейших статистических исследований структурированного и размеченного таким образом текста, то подобные исследования могут и должны строиться с учетом возможности детализации видов прямой речи или же без нее исходя из конкретных филологических задач.

Это, в частности, означает, что принятая нами схема детализации видов прямой речи не может считаться окончательной : уточнение видов речи и соответствующая этим видам маркировка абзацев и предложений будут определяться конкретными потребностями в проведении филологических исследований. Однако и сейчас диалоговая система корректировки текстов *'DIAKOR'* (см.ниже) позволяет, при необходимости, вводить и использовать новые маркеры для абзацев и предложений структурированных текстов, например, выделять косвенную, полупрямую и несобственно-прямую речь.

Диалоговая система корректировки структурированного текста. Система *'DIAKOR'* разработана как автоматизированное рабочее место филолога для выполнения работ по корректировке структурированных прозаических текстов. Необходимость в разработке такой системы диктовалась тем обстоятельством, что для выполнения подобных филологических работ нужна простая, доступная любому пользователю-непрофессионалу компьютерная система; что же касается известных универсальных диалоговых систем *'ЭЕС'*, *'PRIMUS'*, 'ОКО' и др.(Зверев В.И. и др.,1986), то они не только сложны для непрофессионала, но и не обеспе-

чивают выполнения ряда специальных операций, например перенумерации предложений структурированного текста. Система 'DIAKOR' реализована в среде ОС ЕС и предоставляет пользователю возможность работать в режимах чтения, записи, удаления и корректировки (включая перенумерацию) предложений, а также в режиме помощи 'HELP' для получения пользователем основных навыков диалога с ЭВМ.

Предложения структурированного текста выступают в этой системе как квант информационного потока при обмене сообщениями между пользователем и ЭВМ. Организация такого обмена производится следующими основными командами :

1. RR - читать предложение на экран дисплея;
2. WW - писать предложение в память ЭВМ;
3. DD - удалить предложение;
4. RN - перенумеровать текст;
5. HH - перейти в режим 'HELP'.

Выбор автора повествования и настройка системы на определенный текст этого автора производятся пользователем в режиме 'меню'. Вначале на экране дисплея появляется 'СПРАВКА ОБ АВТОРАХ', например :

```
==================================================
*      С П Р А В К А   ОБ   А В Т О Р А Х        *
*              = 0 - конец работы                  *
*              = 1 - Лермонтов М.Ю.                *
*              = 2 - Чехов А.П.                    *
*              = 3 - Толстой Л.Н.                  *
*      Внимание! Для выбора автора замените        *
*          = на * и нажмите на < ВВОД >            *
==================================================
```

После выбора одного из авторов на экране появляется 'СПРАВКА О ТЕКСТАХ', в которой указаны хранящиеся в ЭВМ структурированные тексты этого автора, например :

```
==================================================
*  ЛЕРМОНТОВ М.Ю.                                  *
*       С П Р А В К А   О   Т Е К С Т А Х          *
*  = 0 - конец работы с текстами данного автора    *
*  = 1 - Герой нашего времени                      *
*  = 2 - Княгиня Лиговская                         *
*  = 3 - Панорама Москвы                           *
*     Внимание! Для выбора текста замените         *
*          = на * и нажмите на < ВВОД >            *
==================================================
```

Далее система переходит в основной режим - режим чтения и корректировки структурированного текста.

Отличительной особенностью системы *'DIAKOR'* является возможность работы с записями переменной длины индексно-последовательных файлов. При структуризации текста каждое вычленяемое предложение получает свой номер (ключ) записи, а длина самой записи варьируется в зависимости от длины предложения и может достигать 6000 символов. Для работы с такими записями потребовалось разработать специальные многоэкранные средства связи и поддержки диалога. Наиболее сложным в системе *'DIAKOR'* является режим перенумерации предложений. Действительно, каждое предложение структурированного текста записывается в память ЭВМ под определенным номером (ключом), который соответствует порядковому номеру предложения от начала текста. Поэтому при удалении или вставке нового предложения вся последовательность номеров предложений (начиная с данного номера) должна быть изменена; это требует автоматического слияния двух файлов (основного файла и файла корректуры) с одновременной перенумерацией записей (предложений). Для пользователя, однако, этот режим ничем от других режимов не отличается и по его завершении пользователь получает сообщение о возможности продолжать работу.

В Приложении приведен отрывок романа М.Ю.Лермонтова 'Герой нашего времени' (т.е. отрывок текста, введенного в ЭВМ в виде, идентичном оригиналу) и соответствующий ему фрагмент структурированного текста.

Выводы. В работе представлен формально-пунктуационный метод структуризации художественной прозы, образующий вместе с системой диалоговой корректировки текстов программно-алгоритмический комплекс подготовки и формирования текстовой части МФРЯ. Работоспособность этого комплекса и его высокая 'лингвистическая' надежность подтверждены при структуризации полного текста романа М.Ю.Лермонтова 'Герой нашего времени'. Положено только самое начало большой филолого-компьютерной работе над художественными текстами : и потому, что методы структуризации художественной прозы, а также их компьютерная реализация могут и будут продолжать совершенствоваться, и потому, что те же проблемы существуют для других видов текстов, и по многим другим причинам (см.,напр,:МФРЯ,1986).Следующий шаг, очевидно, должен и будет состоять в детальной инвентаризации филологических потребностей (Герд А.С.,1986) для углубления начатых исследований в рамках МФРЯ.

ПРИЛОЖЕНИЕ

Отрывок из романа М.Ю.Лермонтова 'Герой нашего времени'

А. Фрагмент неструктурированного текста

- *А ЧТО Х ТАКОЕ ОНА ПРОПЕЛА, НЕ ПОМНИТЕ ЛИ?

- *ДА, КАЖЕТСЯ, ВОТ ТАК : '*СТРОЙНЫ, ДЕСКАТЬ,НАШИ МОЛОДЫЕ ДЖИГИТЫ, И КАФТАНЫ НА НИХ СЕРЕБРОМ ВЫЛОЖЕНЫ, А МОЛОДОЙ РУССКИЙ ОФИЦЕР СТРОЙНЕЕ ИХ, И ГАЛУНЫ НА НЕМ ЗОЛОТЫЕ. *ОН КАК ТОПОЛЬ МЕЖДУ НИМИ; ТОЛЬКО НЕ РАСТИ, НЕ ЦВЕСТИ ЕМУ В НАШЕМ САДУ'. *ПЕЧОРИН ВСТАЛ, ПОКЛОНИЛСЯ ЕЙ, ПРИЛОЖИЛ РУКУ КО ЛБУ И СЕРДЦУ И ПРОСИЛ МЕНЯ ОТВЕЧАТЬ ЕЙ; Я ХОРОШО ЗНАЮ ПО ИХНЕМУ И ПЕРЕВЕЛ ЕГО ОТВЕТ.

*КОГДА ОНА ОТ НАС ОТОШЛА, ТОГДА Я ШЕПНУЛ *ГРИГОРЬЮ *АЛЕКСАНДРОВИЧУ: '*НУ ЧТО, КАКОВА?' - '*ПРЕЛЕСТЬ!-ОТВЕЧАЛ ОН. - *А КАК ЕЕ ЗОВУТ?' - '*ЕЕ ЗОВУТ *БЭЛОЙ', - ОТВЕЧАЛ Я.

*И ТОЧНО, ОНА БЫЛА ХОРОША: ВЫСОКАЯ, ТОНЕНЬКАЯ, ГЛАЗА ЧЕРНЫЕ, КАК У ГОРНОЙ СЕРНЫ, ТАК И ЗАГЛЯДЫВАЛИ К ВАМ В ДУШУ.

Б. Фрагмент структурированного текста

=196= %П% - *А ЧТО Х ТАКОЕ ОНА ПРОПЕЛА, НЕ ПОМНИТЕ ЛИ?

=197= %В% - *ДА, КАЖЕТСЯ, ВОТ ТАК: '*СТРОЙНЫ, ДЕСКАТЬ, НАШИ МОЛОДЫЕ ДЖИГИТЫ, И КАФТАНЫ НА НИХ СЕРЕБРОМ ВЫЛОЖЕНЫ, А МОЛОДОЙ РУССКИЙ ОФИЦЕР СТРОЙНЕЕ ИХ, И ГАЛУНЫ НА НЕМ ЗОЛОТЫЕ.

=198= %в% *ОН КАК ТОПОЛЬ МЕЖДУ НИМИ; ТОЛЬКО НЕ РАСТИ, НЕ ЦВЕСТИ ЕМУ В НАШЕМ САДУ'.

=199= %в% *ПЕЧОРИН ВСТАЛ, ПОКЛОНИЛСЯ ЕЙ, ПРИЛОЖИЛ РУКУ КО ЛБУ И СЕРДЦУ И ПРОСИЛ МЕНЯ ОТВЕЧАТЬ ЕЙ; Я ХОРОШО ЗНАЮ ПО-ИХНЕМУ И ПЕРЕВЕЛ ЕГО ОТВЕТ.

=200= %С% *КОГДА ОНА ОТ НАС ОТОШЛА, ТОГДА Я ШЕПНУЛ *ГРИГОРЬЮ *АЛЕКСАНДРОВИЧУ: '*НУ ЧТО, КАКОВА?'

=201= %с% - '*ПРЕЛЕСТЬ! - ОТВЕЧАЛ ОН.

=202= %с% - '*А КАК ЕЕ ЗОВУТ?'

=203= %с% - '*ЕЕ ЗОВУТ *БЭЛОЙ', - ОТВЕЧАЛ Я.

=204= %А% *И ТОЧНО, ОНА БЫЛА ХОРОША: ВЫСОКАЯ, ТОНЕНЬКАЯ, ГЛАЗА ЧЕРНЫЕ, КАК У ГОРНОЙ СЕРНЫ, ТАК И ЗАГЛЯДЫВАЛИ К ВАМ В ДУШУ.

Л И Т Е Р А Т У Р А

Герд А.С. Русская морфология и Машинный фонд русского языка. – ВЯ,1986,№6,с.90–96.

Гринбаум О.Н. Структуризация художественной прозы с использованием ЭВМ. 1.Формально-пунктуационный метод структуризации. //Учен. зап. Тартуского университета. Вып. 827. Тарту, 1988. – С. 74–88.

Зверев В.И., Кетков Ю.Л., Максимов В.С. Алфавитно – цифровые дисплеи ЕС-7920 в диалоговых системах. М.: Наука, 1986, 240 с.

Кожевникова Н.А. О соотношении речи автора и персонажа. – В кн.Языковые процессы современной русской художественной литературы.Проза.М.:Наука,1977,с.7–98.

Машинный фонд Русского Языка: идеи и суждения.М.:Наука,1986, 240 с.

Шапиро А.Б.Современный русский язык.Пунктуация.М.,1974,287с.

BELLES – LETTRES STRUCTURIZING BY MEANS OF COMPUTER

2.STRUCTURIZED TEXT DETAILING

OLEG N.GRINBAUM

S U M M A R Y

THE ARTICLE DEALS WITH THE FURTHER CONSIDERATION OF THE FORMAL-PUNCTUATION METHOD OF BELLES-LETTRES STRUCTURIZING (GRINBAUM,1987). THE METHOD OF AUTOMATIC MARKING-OUT OF PARAGRAPHS AND SENTENCES ACCORDING TO THE WRITTEN FORM TYPES OF SPEECH IS CONSIDERED. THE ALGORITHM OF DETAILING THE DIRECT SPEECH TYPES IS PRESENTED.THE SYSTEM OF CORRECTION OF A STRUCTURIZED TEXT IN A DIALOGUE IS DESCRIBED. THE MAIN RESULTS OF DETAILING ARE ILLUSTRATED ON THE BASIS OF LERMONTOV'S PROSE.

ГИПЕРЛЕКСЕМНАЯ ГРУППИРОВКА СЛОВ КАК СПОСОБ ПРЕДСТАВЛЕНИЯ СИСТЕМНОСТИ ЛЕКСИКИ

Г.О. Каримова

В современной лексикологии одной из центральных является проблема, связанная с выявлением и изучением системной организации лексики. Осмысление системных отношений в лексике породило различные классификации словарного состава - от семантических полей, где лексические единицы объединяются с учетом самых общих лексико-семантических связей, до традиционных систем словоизменения и формообразования, нетрадиционно рассматриваемых сквозь призму парадигматики (например, внутрисловная парадигма человек - люди, или парадигмы склонения и спряжения). Многообразие классификаций словарного состава показывает возможность выделения самых разнообразных признаков, учитывающих разные стороны лексических единиц и связывающих их в единое целое на основе сходства по одним признакам и противопоставленности по другим. И не случайно в работах последних лет предпринимается попытка построить типологию парадигм (Шмелева Т.В., 1987), очертить логическое пространство возможных единиц лексической системы языка (Поликарпов А.А., 1988).

Путь к наиболее полному и всестороннему описанию лексической системы языка лежит через поиски исследованных ранее группировок лексических единиц. Цель настоящей публикации - сделать несколько шагов в направлении определения лингвистического статуса одной из наименее изученных единиц лексической системы языка - гиперлексемы.

Обратимся сначала к употреблению в литературе термина гиперлексемы (далее - ГЛ), отличающемся довольно развитой полисемией.

Этот термин мы встречаем в работах О.С. Ахмановой (Ахманова О.С., 1967) и Э.М. Медниковой (Медникова Э.М., 1972), где под ГЛ подразумевается лексико-морфологическая категория (например, ГЛ процессности), реализующаяся в ряде аллогиперлекс (например, "категория агентивности с категориальными формами исполнителя и объекта действия" является одной из аллореализаций ГЛ процессности). Понятие ГЛ, таким образом, используется в целях изучения механизма "лексической морфологии".

В.А. Успенский (Успенский В.А., 1977) предложил употреблять термин ГЛ для описания различных спорных случаев лек-

сикографического толкования, возникающих при описании залогов. Так, в предложениях Человек катит колесо и Обруч катится слова-партиципанты колесо и обруч имеют одну и ту же роль, следовательно, катит и катится объединяются в одну ГЛ. И.Б. Долинина, развивая положения, высказанные в упомянутой работе В.А. Успенского, определяет ГЛ как "такое значение глагольной вокабулы, которое включает в себя максимальное количество партиципантов, находящих синтаксическое выражение хотя бы в одной из структур" (Долинина И.Б., 1978, с. 164). Понимание ГЛ В.А. Успенским и И.Б. Долининой тесно связано с целями исследования залога и диатезы.

На совершенно ином участке языковой структуры предлагает использовать термин ГЛ Л.В. Малаховский. Исследуя особенности английской омонимики, он выявляет несколько типов омонимов, среди которых лексико-грамматические омогруппы, или ГЛ, объединяющие слова, "близкие друг другу по лексической семантике, но принадлежащие к разным грамматическим классам (частям речи)" (Малаховский Л.В., 1987, с. 8). Например, light "светлый" (прилагательное), light "свет", "освещение" (существительное), light "освещать" (глагол) образуют одну ГЛ с общим значением "освещенности".

Рассмотренные трактовки ГЛ обнаруживают принципиальное сходство - под ГЛ, как правило, подразумевается некоторая инвариантная единица. И поскольку существует аллоэмический подход к явлениям языка, а понятие "гипер-эмы" прочно вошло в теорию современного языкознания, постольку и область применения этого термина будет постоянно расширяться.

Термин ГЛ в последнее время активно используется и в интересах лексикологии и лексикографии.

В Частотном словаре индексирования Л.В. Сахарного (Сахарный Л.В., 1974) впервые, пожалуй, в отечественной лексикографии ГЛ избирается в качестве основной словарной единицы, в которой отождествляется ряд однокоренных слов, связанных отношениями трансформации. Например, исходная лексема достаточный, трансформируясь в различные грамматические категории (достаточен - достаточно - достаточность) сохраняет тождество лексических значений и образует отадъективную ГЛ. Аналогичные отношения трансформации наблюдаются и между лексемами в составе отглагольной ГЛ включать - включаться - включающий - включаемый - включая - включенный - включительно - включение - включен.

Выбор этой единицы кодирования смыслового содержания текста представляется убедительным, так как в ней оказывается снятой категориально-грамматическая вариативность лексем, не имеющая отношения к денотативному аспекту текста, что и интересует нас при индексировании и поиске текстов в информационно-поисковых системах.

П.Н. Денисов, рассуждая о разнообразии классификаций, существующих в лексике, выделяет и "классификацию по словообразовательным гнездам (гиперлексемам), в которых в одно гнездо попадают слова, принадлежащие к разным частям речи на основе смысловой близости (читать, чтение, чтиво, читатель, чтец, читальня, читальный, читающий)" (Денисов П.М., 1976, с. 64). В этом определении не конкретизирована степень смысловой близости между словами, объединенными в одну ГЛ (ср. семантическое расстояние между читать и читающий, с одной стороны, и читать и читальня, - с другой), и поэтому не очерчена граница между ГЛ и словообразовательным гнездом (далее - СГ), которые, скорее всего, автор склонен отождествлять. В то время как в иерархическом строении лексической подсистемы языка каждая из этих единиц занимает свое место (Поликарпов А.А., 1988).

И ГЛ и СГ, как и любые другие единицы определенного уровня языка, выявляются на основе сходства и различия по тем или иным признакам плана выражения и плана содержания. Снятие признака категориально-грамматического варьирования и отождествление словообразовательно связанных слов с точностью до тождества корневой части плана выражения и корневой семантики в плане содержания приводит к образованию такой единицы как СГ. В ГЛ отождествление словообразовательно связанных слов проходит по линии границы производящей основы, а в плане содержания обязательным является тождество собственно лексических компонентов хотя бы в одном из значений. Так, например, лексемы забегать и забегаловка - члены одного СГ, но представители разных ГЛ, так как в значении существительного ("маленькая второразрядная закусочная с продажей вина") отсутствуют лексические компоненты, соответствующие значению производящего глагола (ср. значение глагола забегать "Бегом входить, попадать куда-либо //Разг. Заходить куда-либо на короткое время, обычно по пути, мимоходом"). Лексическое значение производного в данном случае (с точки зрения синхронного словообразования) не представляет собой простую сумму значений его частей, а смысловое

приращение, получаемое в результате присоединения словообразовательного средства -ловк- к производящей основе забега- нерегулярно и, следовательно, уникально.

Лексемы же бегать и беганье мы группируем в одну ГЛ и по форме и по содержанию - производное здесь выражает то же лексическое значение, что и производящее, только в оболочке другой части речи.

Если идти дальше, то можно объединить в одну ГЛ наряду с лексемами бегать и беганье еще и лексемы побегать и бегун. Лексическое значение производного побегать описывается в словаре с помощью его мотивирующего бегать с добавлением лексического компонента, конкретизирующего характер совершения действия во времени ("бегать некоторое время"). Аналогичные формально-семантические отношения между словообразовательно связанными словами типа бегать - побегать ("действие, названное мотивирующим глаголом, совершить в течение некоторого времени (чаще недолгого)") наблюдаются и в ряде других случаев: беседовать - побеседовать, заниматься - позаниматься, работать - поработать и т.д. Производные в перечисленных парах глаголов выражают то же основное значение, что и исходные слова, но только несколько видоизменяя и дополняя его.

В отличие от бегать - беганье, бегать - побегать в паре бегать - бегун наблюдается радикальное изменение производного по сравнению с исходным словом и в отношении перехода производного в иную грамматическую категорию и в отношении его лексического значения. В данном случае основанием для включения лексемы бегун в рассматриваемую ГЛ является регулярное появление соответствующего семантического компонента в серии производных с аналогичным суффиксом (ср.: бегун - "тот, кто бегает", опекун - "тот, кто опекает", прыгун - "тот, кто прыгает" и т.п.).

Итак, мы объединяем в одну ГЛ лексемы бегать, беганье, побегать, бегун. Будучи словообразовательно связанными, они обладают сходством в плане выражения. В плане содержания они характеризуются либо тождеством лексических компонентов хотя бы в одном из значений (бегать - беганье), либо регулярностью появления соответствующего значения в ряде слов с аналогичным словообразовательным средством (бегать - бегун). Причем, совсем не обязательно, чтобы во всех членах гиперлексемы наличествовало сквозное тождество по какому-либо

общему для всех них значению. Так, например, лексемы бегать, побегать, бегун только в оппозиции к лексеме бегать оказываются в таких формально-семантических отношениях, которые удовлетворяют условиям для включения их в одну ГЛ. В эту же ГЛ входит и лексема бегунья (и ряд других лексем), которая в сопоставлении с лексемой бегун обнаруживает требуемое формально-семантическое тождество (суффикс -j- здесь сопровождается регулярным изменением значения слова в сторону приписывания ему компонента "женскости"; ср.: прыгун - прыгунья, шалун - шалунья и т.п.).

Принципы объединения лексем в ГЛ (в особенности при выяснении степени преобразования исходных базовых лексем в актах словообразования) можно в самом общем виде свести к трем типам словопроизводства - синтаксической деривации, модификационной лексической деривации и мутационной лексической деривации.

При синтаксической деривации лексическое значение производного частично или полностью воспроизводит лексические значения производящего, за исключением категориально-грамматических компонентов, т.е. принадлежности слова к той или иной части речи (бегать - беганье, бегать - беговой).

При лексической деривации в семантике производного слова появляются такие лексические семантические компоненты, которые отсутствуют в семантике производящего. Однако в пределах одной ГЛ могут объединяться и лексические дериваты при условии, если добавочные лексические компоненты, появляющиеся в их значении, регулярно связаны с употреблением определенного словообразовательного средства. Одним из типов лексической деривации с регулярным семантическим результатом является модификация. Сущность ее заключается в добавлении к основному значению мотивирующего слова некоторого дополнительного признака: выделяются модификационные значения женскости (бегунья, беженка), собирательности (беженство), невзрослости, подобия, единичности и другие. Как и при синтаксической деривации, в модификационных образованиях обнаруживается регулярная повторяемость формальных и семантических отношений словообразовательно связанных слов.

Мутация наблюдается при лексической деривации, когда производное означает иное понятие по сравнению с тем, которое выражается производящим словом. Но и мутационные образования в некоторых случаях представляют собой относительно

регулярные конструкции (типа бегать - бегун) и, следовательно, должны быть включены в одну ГЛ наряду с синтаксическими дериватами и модификационными лексическими дериватами.

Если при мутации в смысловой структуре производного появляется уникальный добавочный семантический компонент, идиоматически, т.е. неповторимо, связанный с производящим средством, а происходящие в нем изменения нельзя описать достаточно общими правилами, то данная производная лексема рассматривается как отдельная ГЛ. Так, например, существительное взятка образовано от глагола взять путем прибавления суффикса -к-. Лексическое значение производного не представляет собой простую сумму его частей. Если суммировать значение частей слова взятка (значение производящего глагола взять и значение отвлеченного процессуального признака суффикса -к-), то можно получить лишь значение "то, что взято", но это не будет лексическим значением данного слова: оно обозначает не просто "то, что взято", а - "Подарок, вознаграждение должностному лицу с целью заставить его сделать что-либо в интересах дающего; вынужденные поборы с зависимых и подчиненных представителями власти". В семантической структуре данного производного смысловое приращение занимает весь объем его лексического значения. Более того, смысловое приращение в значении слова взятка не наблюдается больше ни в каких других отглагольных существительных с данным суффиксом. Производное взятка относится к числу мутационных лексических дериватов с нерегулярной формально-семантической воспроизводимостью и, следовательно, образует отдельную ГЛ, в которую наряду с ним будут входить лексемы взяточнический, взяточничать (синтаксические дериваты), взяточник (регулярный лексический дериват), взяточница, взяточничество (лексические дериваты с модификационным значением "женскости" и "собирательности"). На периферии этой ГЛ будет располагаться устаревшее существительное взяток ("то же, что взятка").

Что касается исходной лексемы взять, то она вместе с лексемами взяться и взятие образует отдельную ГЛ. Лексема взятие является синтаксическим дериватом глагола взять и воспроизводит в своей семантической структуре три его значения (I. Овладевание, завоевание; захват. (взятие Измаила - взять Измаил). 2. Задержание, арест. (взятие под стражу -

взять под стражу). 3. Получение чего-либо в свое обладание, пользование.)) Покупка, приобретение чего-либо. (деньги на взятие места на пароходе - деньги для того, чтобы взять место на пароходе). Взяться - страдательная форма глагола взять по первому значению (взяться за блюдечко - взять блюдечко).

Таким образом, граница между членами разных ГЛ, по нашим наблюдениям, проходит по линии, противопоставляющей синтаксическую деривацию и лексическую деривацию с регулярными изменениями в плане выражения и плане содержания с одной стороны, и лексическую деривацию, не сопровождаемую регулярными изменениями в плане выражения и плане содержания, с другой.

В организации семантической структуры ГЛ принимают участие все ее лексемные составляющие. Так, у ГЛ, возглавляемой лексемой взять, 15 значений. Подавляющая часть их принадлежит глаголу взять (12 значений). У лексемы взяться 4 значения, по одному из них она соотносится с первым значением исходного глагола, а три других, являясь специфичными значениями лексемы взяться, дополняют семантический объем ГЛ. Синтаксический дериват взятие воспроизводит в своей семантической структуре 2-ое, 4-ое, 6-ое значения исходного глагола. Семантическую связь между лексемами в составе рассматриваемой ГЛ можно назвать радиальной: в центре - исходная лексема взять, от которой на приблизительно одинаковом смысловом расстоянии располагаются семантически не связанные между собой лексемы взяться и взятие.

Между членами ГЛ возможны и другие типы семантической связи. Так, например, сквозная связь, когда все лексемы в ГЛ имеют одно общее значение - взятка, взяточник, взяточница, взяточничать, взяточничество, взяток связываются по первому значению лексемы взятка. Наглядно отношения между семантическими структурами указанных слов в пределах ГЛ "взять" и "взятка" представлены ниже на рис. 1а и рис. 1б.

При цепочечной связи каждая пара слов связывается по какому-то значению, не участвующему в подобных отношениях каждого из слов этой пары с другими членами ГЛ.

ГЛ "взять"

Лексемы с их отсылкой к сводной семантической структуре ГЛ	Сводная семантическая структура ГЛ I 2 3 4 5 6 7 8 9 I0 II I2 I3 I4 I5
взять	
I	+
2	+
3	+
4	+
5	+
6	+
7	+
8	+
9	+
I0	+
II	+
I2	+
взяться	
I	+
2	+
3	+
4	+
взятие	
I	+
2	+
3	+

рис. Iа

ГЛ "взятка"

Лексемы с их отсылкой к сводной семантической структуре ГЛ	Сводная семантическая структура ГЛ I 2 3
взятка	
I	+
2	+
3	+
взяток	
I	+
2	+
взяточник	
I	+
взяточница	
I	+
взяточничество	
I	+
взяточнический	
I	+
взяточничать	
I	+

рис. Iб

Звеном, связывающим в одну ГЛ лексемы бегать и беглость, выступает наличие тождественных лексических компонентов в одном из значений исходного глагола бегать ("совершать побег") и его производного беглый ("совершивший побег откуда-либо"), синтаксическим дериватом которого является лексема беглость. Причем, последняя воспроизводит в своей семантической структуре те значения, по которым лексема беглый не соотносится ни с какими другими членами ГЛ (за исключением наречия бегло). Более наглядно семантическую связь между лексемами бегать, беглый и беглость можно представить следующим образом: (см. рис. 2).

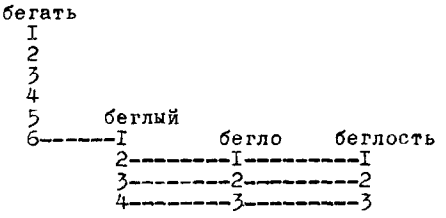

рис. 2

Мы полагаем, что центральной единицей лексической подсистемы языка является ГЛ.[+] В качестве объективного основания для ее выделения выступает и единство формальной структуры и лексико-семантическая близость слов. Отождествление в ГЛ ряда однокоренных слов, обладающих единством индивидуально-лексических сем, позволяет ей сохранять инвариантный план лексического выражения и инвариантный план лексического содержания.[++]

ЛИТЕРАТУРА

Ахманова О.С. Некоторые особенности глагольной гиперлексемы в русском языке // "To Honor Roman Jakobson"; v. I, - The Hague-Paris, 1967, с. 141-149.

Денисов П.Н. Системность и связность в лексике и система словарей. // Проблематика определений терминов в словарях разных типов. Л., 1976, с. 63-73.

[+] О пространстве единиц лексической системы языка см. (Поликарпов, 1988; Каримова, Поликарпов, 1989).

[++] Хотелось бы подчеркнуть, что наше представление о ГЛ созвучно точке зрения Л.В. Сахарного. Особенно в отношении практической необходимости выделения этой единицы, в пределах которой "могут быть выяснены широкие чисто семантические характеристики информативности, не связанные с категориальными отношениями, характерными для лексем, а тем более для словоформ". (Сахарный, 1974, с. 8).

Долинина И.Б. Рефлексив и средний залог в системе английских залогов и проблема "гиперлексемы" // Проблемы теории грамматического залога. Л., 1978, с. 162-171.
Каримова Г.О., Поликарпов А.А. Принципы выделения гиперлексемы как единицы лексической системы языка. (в печати)
Малаховский Л.В. Структура английской омонимики и ее отражение в словарях // ИЯШ, 1987, № 2, с. 8-12.
Медникова Э.М. Некоторые особенности английской гиперлексемы // Актуальные проблемы лексикологии и лексикографии. Пермь, 1972, с. 337-341.
Поликарпов А.А. Логическое пространство единиц лексической подсистемы языка и квантификация соотношений между ними // Прикладная лингвистика и автоматический анализ текста. Тез. докладов научной конференции 28.01 - 30.01 1988 г. Тарту, 1988, с. 67-69.
Сахарный Л.В. Частотный словарь индексирования. Пермь, 1974.
Успенский Б.А. К понятию диатезы // Проблемы лингвистической типологии и структуры языка. Л., 1977, с. 65-84.
Шмелева Т.В. (ред.) Системный анализ значимых единиц русского языка. Парадигматика в лексике и словообразовании. Красноярск, 1987. - 144 с.

HYPERLEXEMIC GROUPING OF WORDS AS A WAY OF REPRESENTING THE LEXICAL SYSTEM

Gulmira Karimova

S u m m a r y

The article deals with the determination of the linguistic status of hyperlexeme. The HL preserves the invariant plane of lexical expression, an invariant plane of lexical content, because in it one-stem words possessing the same lexical meaning are converged.

ЧАСТОТНО-РАСПРЕДЕЛИТЕЛЬНЫЙ СЛОВАРЬ ОТДЕЛЬНЫХ РАССКАЗОВ В.М. ШУКШИНА

А.И.Крюченков

Материалом для формирования выборки и составления частотно-распределительного словаря рассказов В.М.Шукшина послужили тексты 57 рассказов названного автора,что составляет около 40% от общего числа его рассказов.

Общая длина обследованных текстов - II7729 словоупотреблений. Для изучения взяты тексты различной длины - от 904 до 4I73 словоупотреблений.Материал взят из различных по тематике сборников рассказов В.М.Шукшина:"Сельские жители","Беседы при ясной луне","Характеры". Тексты обрабатывались от начала и до конца.

Частотно-распределительный словарь содержит словоформы с частотой от I и выше,всего 50I6 еденицы (без учета словоформ с частотой I). Словарь составлен с опорой на словоформы. Они расположены в порядке убывания частот. Наибольшую частоту имеет "и" - 3688. Для каждой единицы словаря указаны:её ранг(порядковый номер по убыванию частоты),абсолютная частота и количество текстов,в которых встретилась данная словоформа (политекстия). Кроме того,даётся частота слова в каждом из 57 текстов. Нумерация текстов от I до 57 произвольная. В предложенной таблице приводятся I00 словоформ,имеющих наивысшие показатели абсолютной частоты.

Тексты рассказов в соответствии с нумерацией и с указанием количества словоупотреблений располагаются следующим образом:

I. "Дядя Ермолай" - I039 словоупотреблений
2. "В воскресенье мать старушка" - I360
3. "На кладбище" - I463
4. "Горе" - I3I3
5. "Как помирал старик" - 999
6. "Сильные идут дальше" - I87I
7. "Билетик на второй сеанс" - I957
8. "Привет Сивому" - I956
9. "Раскас" -I703
I0. "Письмо" - I577
II. "Два письма" - I583

I2. "Экзамен" - I74I
I3. "Версия" - I874
I4. "Петька Краснов рассказывает" - I2I9
I5. "Материнское сердце" - 3645
I6. "Чудик" - 2354
I7. "Алеша Бесконвойный" - 4I73
I8. "Миль пардон,мадам" - 2022
I9. "Залётный" - I958
20. "И разыгрались же в поле кони" - 2067
2I. "Срезал" - 2I27
22. "Как зайка летал на воздушных шариках" - 42I8
23. "Наказ" - 3268
24. "Мой зять украл машину дров" - 3500
25. "Шире шаг,маэстро" - 3367
26. "Нечаянный выстрел" - I904
27. "Волки" - I636
28. "Бессовестные" - 2843
29. "Хахаль" - 2373
30. "Стёпка" - 2673
3I. "Микроскоп" - 2266
32. "Земляки" - 2352
33. "Мастер" - 2894
34. "Петя" - I3I2
35. "Забуксовал" - I2I9
36. "Операция Ефима Пьяных" - I400
37. "Хозяин бани и огорода" - I529
38. "В профиль и анфас" - 2388
39. "Сапожки" - I882
40. "Думы" - I602
4I. "Мнение" - I465
42. "Мужик Дерябин" - 904
43."Рыжий" - I228
44. "Ораторский приём" - I605
45. "Ноль-ноль целых" - I085
46. "Выбираю деревню на жительство" - 2259
47. "Обида" - 2II5

48. "Беспалый" - 2783
49. "Космос,нервная система и шмат сала" - 2579
50. "Крепктй мужик" - 1553
51. "Вянет,пропадает" - 1373
52. "Одни" - 1635
53. "Генерал Малафейкин" - 2013
54. "Сельские жители" - 2126
55. "Капроновая елочка" - 2979
56. "Пьедестал" - 2893
57. "Психопат" - 2503

В словнике обращают на себя внимание отдельные слова, имеющие высокий ранг при малом количестве текстов,в которых данное слово встречается. Это имена собственные. В первой сотне - это Иван (абсолютная частота 121 в 12 текстах). Далее идут собственные имена со следующими показателями: Федор - а.ч. 108 в 2-х текстах,Алёша - а.ч. 90 в 1 тексте,Павел - а.ч. 76 в 1 тексте,Сашка - а.ч. 62 в 2-х текстах. Такие знаменательные слова в рассказах часто заменяются нарицательными словами,имеющими ту же функцию,что и имена собственные. Они также резко выделяются в словнике,например: старик - а.ч. 112 в 10 текстах, мать - а.ч. 157 в 30 текстах,кандидат - а.ч. 55 в 2 текстах, чудик - а.ч. 50 в 1 тексте и т.д. Такого рода замены нужно иметь в виду при составлении общей выборки.

Ограниченный объем статьи не позволяет проводить детальный анализ,поэтому здесь приводится в основном только материал справочного характера,анализ же будет проведен в последующих работах.

Ранг		Абсол. част.	Поли-текст.	№ текстов																										
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	и	3688	57	31	53	63	32	18	63	67	71	61	35	42	48	55	30	94	56	182	43	70	62	69	135	115	105	107	71	50	97	86
2	не	3050	57	33	32	34	37	31	38	60	47	54	53	41	39	44	36	98	74	104	42	58	48	62	100	113	88	75	36	34	88	65
3	в	2610	57	25	29	20	21	10	51	32	42	53	25	50	45	50	18	81	58	112	45	29	55	37	84	47	91	94	59	48	50	46
4	на	1993	57	25	25	29	12	19	30	30	24	36	18	25	31	41	28	46	39	67	43	26	52	50	72	50	56	54	40	27	43	35
5	а	1744	57	14	22	31	33	21	16	35	14	23	42	31	9	20	20	55	46	42	30	15	22	26	68	59	57	31	17	10	64	50
6	что	1502	57	14	12	13	3	2	27	20	38	19	12	23	39	23	8	57	26	48	12	25	23	43	55	46	50	52	10	13	33	39
7	он	1441	57	9	19	9	11	22	30	16	37	30	5	10	20	18	11	46	34	66	25	26	27	24	43	48	54	51	44	17	34	28
8	я	1188	57	19	8	28	26	8	10	26	14	40	22	23	25	31	13	20	20	12	34	22	29	11	55	39	33	30	10	13	35	23
9	с	1179	57	13	13	12	18	10	16	24	19	14	14	16	10	16	9	36	20	28	22	12	12	20	43	30	31	39	21	26	31	28
10	как	1021	57	13	7	16	6	5	11	16	19	14	6	22	23	24	13	24	18	32	20	12	26	22	44	45	42	19	20	4	31	25
11	так	882	57	2	4	21	9	7	9	17	7	12	10	10	8	16	10	19	7	45	9	16	9	22	43	26	27	13	12	7	24	17
12	это	772	57	3	9	11	5	3	13	9	20	22	9	7	27	12	7	18	14	26	17	19	19	15	18	30	24	23	6	7	11	16
13	у	765	57	4	7	5	10	2	5	10	16	13	19	19	8	10	7	27	16	20	13	12	13	6	30	20	21	21	7	6	22	23
14	же	729	57	7	7	15	2	4	8	6	13	14	11	4	7	10	5	29	17	24	4	13	8	9	22	26	22	22	4	10	21	9
15	ты	723	53	3	9	15	7	8	8	20	10	-	28	13	-	13	8	25	8	9	7	9	18	2	27	25	32	-	7	21	23	5
16	да	675	56	12	2	15	8	7	8	11	4	7	20	11	11	9	10	30	13	22	11	10	8	9	45	20	13	15	2	9	16	14
17	но	658	57	5	7	5	3	5	14	9	15	19	10	8	8	12	5	19	9	41	7	12	12	21	25	34	24	28	7	5	17	11
18	всё	627	57	8	6	12	11	2	8	10	9	12	12	12	7	5	4	26	9	30	8	24	8	7	24	25	21	17	6	8	24	6
19	к	586	57	3	5	7	4	2	13	15	7	6	5	6	6	7	3	15	8	17	8	7	11	17	17	19	9	26	8	9	20	21
20	она	558	50	-	5	23	2	5	I	8	11	24	10	I	-	15	7	31	14	13	I	13	3	2	15	6	57	9	3	-	24	10
21	ну	548	55	7	3	12	6	4	8	13	9	6	3	8	5	12	18	18	6	18	9	11	4	8	31	16	16	8	1	11	14	19
22	его	536	57	2	9	7	4	4	7	3	13	7	4	2	4	8	3	20	19	22	15	16	9	6	15	24	21	11	7	10	5	6
23	вот	525	56	3	5	15	8	2	8	23	-	7	8	9	6	4	5	19	13	25	8	4	5	6	26	33	14	19	5	4	20	14
24	за	501	56	3	4	1	-	2	9	7	7	5	12	6	5	8	6	18	14	17	12	5	12	4	17	12	26	12	10	8	15	14
25	по	476	57	6	4	5	6	2	13	11	3	5	2	3	5	6	6	18	10	15	11	4	15	10	9	16	9	17	8	3	10	7
26	ещё	448	56	1	2	7	-	2	9	6	11	6	4	6	7	15	5	13	11	34	9	13	6	5	23	5	9	8	7	9	5	9
27	сказал	408	56	1	5	1	1	9	3	6	12	4	1	1	5	17	4	2	10	5	2	8	8	8	31	9	9	10	9	4	-	4
28	надо	396	56	-	4	6	3	3	5	4	5	3	4	3	5	2	-	8	13	20	4	11	4	10	16	15	16	23	5	5	13
29	то	395	56	2	2	8	6	5	8	9	8	6	6	11	2	6	8	6	6	12	7	7	7	11	21	12	15	8	13	4	7	9
30	бы	393	55	5	3	4	8	2	5	26	3	8	8	6	5	3	1	13	7	16	4	6	3	6	23	13	16	5	12	5	21	7
31	тоже	387	56	5	3	4	6	4	2	6	9	12	7	3	3	4	2	16	13	13	3	7	5	12	15	16	13	8	3	7	10	8
32	было	374	56	8	6	5	7	4	4	6	14	5	3	4	5	5	2	13	11	10	12	7	6	6	8	11	11	12	3	4	8	10
33	из	359	56	3	5	1	2	1	3	1	2	4	-	2	8	11	-	9	9	16	4	4	7	6	15	8	15	9	6	4	3	4
34	там	358	54	9	3	16	7	4	5	5	8	6	6	2	9	4	16	17	6	13	2	5	6	4	14	10	9	7	6	4	1	9
35	тут	343	56	5	4	12	7	-	3	8	12	6	5	3	4	7	10	12	9	12	6	1	11	13	14	10	15	12	2	2	3	8

Ранг		№ текстов											
		30	31	32	33	34	35	36	37	38	39	40	41
I	и	91	59	62	90	41	37	26	33	80	43	71	56
2	не	49	55	56	70	33	39	47	55	69	52	39	28
3	в	72	52	63	77	32	22	38	28	43	35	38	28
4	на	46	39	46	44	8	9	24	28	59	36	22	27
5	а	53	26	35	30	20	34	21	39	46	26	29	13
6	что	29	12	13	28	14	19	28	23	16	35	20	26
7	он	28	20	25	43	11	11	14	10	29	21	20	17
8	я	27	19	3	31	18	11	16	38	47	14	7	3
9	с	28	33	28	31	12	12	11	7	28	12	13	18
10	как	11	15	14	30	6	15	19	16	21	10	14	13
11	так	20	9	12	18	9	25	7	15	19	13	16	9
12	это	10	6	7	18	6	10	6	8	8	17	8	15
13	у	15	24	13	21	5	1	12	13	16	15	9	6
14	же	9	12	10	16	2	20	7	24	11	18	15	8
15	ты	20	30	5	13	17	6	17	18	14	10	11	5
16	да	10	18	17	13	2	14	15	16	16	16	7	11
17	но	12	4	7	16	4	5	8	9	12	8	11	11
18	всё	9	6	14	11	8	11	2	5	16	13	13	9
19	к	30	14	12	16	8	4	8	5	14	8	7	6
20	она	12	14	7	16	11	2	3	-	7	20	8	5
21	ну	10	11	8	8	5	11	9	13	22	6	8	6
22	его	12	7	9	9	3	2	2	12	8	2	4	5
23	вот	12	8	13	9	4	14	8	4	10	5	7	2
24	за	13	11	11	5	3	5	6	8	9	11	8	7
25	по	14	16	10	10	10	7	3	4	9	11	9	18
26	ещё	12	3	15	11	8	2	5	6	3	7	9	6
27	сказал	12	3	13	9	1	3	4	6	15	3	4	12
28	надо	8	8	7	6	4	9	3	7	15	7	7	1
29	то	6	7	2	12	6	4	2	4	8	3	3	3
30	бы	4	8	7	13	2	2	10	10	2	7	2	8
31	тоже	8	6	11	11	1	2	2	8	8	3	4	2
32	было	8	6	10	7	1	2	6	2	9	5	11	4
33	из	8	7	8	14	5	2	4	5	3	3	7	6
34	там	13	5	7	9	3	1	7	4	4	5	1	4
35	тут	5	8	7	2	2	7	4	6	5	4	3	1

Ранг		№ текстов											
		42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57
I	и	37	55	46	30	81	57	116	66	46	33	50	47	61	66	121	76
2	не	8	25	41	33	69	62	69	60	40	30	41	53	44	67	87	63
3	в	31	22	54	17	56	52	48	57	27	26	32	48	38	83	65	49
4	на	16	23	37	26	25	32	45	50	19	24	27	33	42	45	50	38
5	а	15	12	16	18	39	22	33	52	14	19	27	24	35	47	24	52
6	что	13	18	20	28	34	58	52	23	8	6	19	29	29	28	39	50
7	он	17	14	21	11	32	31	46	34	9	11	8	12	19	28	34	51
8	я	6	30	9	13	27	27	18	9	8	6	18	26	20	25	13	20
9	с	13	15	24	7	20	28	32	22	21	13	14	28	20	31	24	31
10	как	12	10	6	10	13	16	21	18	12	13	10	25	18	23	22	29
11	так	6	8	6	8	17	12	28	15	5	8	14	8	30	17	14	20
12	это	9	7	11	5	45	6	36	26	10	7	7	21	2	9	21	16
13	у	8	6	4	5	26	14	17	31	5	12	10	22	12	17	13	22
14	же	7	6	5	10	11	15	34	14	9	4	9	17	14	18	18	33
15	ты	1	-	6	17	14	13	11	24	10	6	16	6	13	20	33	7
16	да	6	-	2	5	15	15	19	5	9	2	3	18	10	5	11	23
17	но	3	10	11	6	14	12	17	16	9	2	7	9	13	4	2	19
18	всё	4	5	3	7	12	9	19	13	9	5	1	5	10	16	19	12
19	к	4	8	9	1	8	11	12	6	10	2	3	15	13	29	9	13
20	она	-	-	3	-	7	12	29	7	5	2	17	1	12	5	26	14
21	ну	3	-	-	1	20	8	10	10	2	8	2	11	12	6	18	16
22	его	11	10	11	3	11	16	21	10	2	1	4	12	9	11	28	18
23	вот	8	5	3	1	12	3	17	15	4	9	4	5	5	5	8	7
24	за	3	3	7	18	8	14	16	8	3	9	10	7	4	6	13	4
25	по	4	3	10	6	6	4	9	11	8	5	4	3	8	12	11	12
26	ещё	3	4	2	5	6	9	10	12	5	9	5	4	14	11	16	16
27	сказал	4	3	8	6	1	4	4	4	3	10	5	14	9	39	9	11
28	надо	3	2	1	2	4	10	8	11	1	8	3	2	6	15	11	11
29	то	-	3	6	7	7	6	12	13	6	1	5	2	12	13	5	5
30	бы	-	4	1	-	5	8	11	9	5	5	10	6	6	3	6	5
31	тоже	-	7	4	2	17	4	5	9	1	5	5	9	9	9	9	7
32	было	-	7	4	1	6	8	8	14	3	2	6	6	4	13	7	4
33	из	3	6	12	1	13	10	4	8	9	4	4	12	7	7	11	13
34	там	-	3	-	1	17	8	4	6	4	-	5	7	7	8	8	4
35	тут	1	3	3	3	6	7	9	1	3	2	2	1	9	2	11	10

Ранг	Словоформа	Абсол. част.	Полн. текст.	№ текстов																										
				I	2	3	4	5	6	7	8	9	10	II	I2	I3	I4	I5	I6	I7	I8	I9	20	2I	22	23	24	25	26	27	28	29
36	чело	34I	54	-	3	I0	5	-	3	I6	3	I	I2	1	1	6	6	9	2	II	I	2	2	2	I8	5	8	7	6	6	I5	5
37	они	329	56	4	5	4	2	I	3	-	4	6	8	I	2	4	5	I7	7	I7	2	0	5	3	4	I6	7	I2	-	5	7	I3
38	нет	328	55	2	4	7	I	3	4	I	3	6	I0	3	9	4	4	5	4	9	3	4	I	6	I8	9	4	I9	7	5	7	II
39	ут	309	54	2	4	-	3	5	5	4	2	4	2	6	3	5	4	7	5	I2	3	8	6	I0	II	7	I0	7	6	7	3	4
40	только	307	56	4	7	6	3	-	7	3	I	4	3	4	3	2	6	II	6	I4	3	I4	5	4	5	8	7	7	6	4	5	II
41	её	295	5I	4	5	5	I	2	3	4	4	5	7	I	I	5	2	I2	7	I2	I	6	3	-	I8	-	3	9	6	-	I5	I0
42	мне	29I	5I	2	5	I6	7	5	3	7	2	5	6	9	II	5	6	7	I	6	3	II	4	I	I0	-	8	I0	2	-	I3	7
43	опять	28I	56	6	4	8	3	I	7	4	8	2	3	5	4	6	2	8	7	9	8	3	4	5	7	9	7	I4	2	I	3	6
44	меня	280	56	I	5	7	3	6	-	6	4	5	I0	I3	6	5	3	7	6	I	9	9	4	I	I0	4	8	6	5	7	9	7
45	когда	274	52	7	2	4	I	I	I	5	3	3	I	2	3	-	-	I2	4	I9	8	5	5	I0	I5	I	II	7	8	-	5	2
46	ль	269	49	9	8	I	-	I	6	2	9	6	7	-	II	4	-	8	7	3	5	8	4	I0	5	3	I0	I7	-	I	I0	5
47	был	267	52	5	6	I	3	2	3	7	I0	4	4	-	7	3	5	8	2	8	I0	0	4	4	I2	9	9	3	7	8	8	4
48	ук	260	55	2	2	4	6	7	I	2	2	3	5	4	I	2	3	I6	2	I6	2	4	6	5	II	9	7	3	4	-	II	4
49	мы	253	5I	I4	2	2	3	-	I	-	6	4	2	II	I	6	-	9	2	3	7	5	2	I3	8	2I	7	4	I	4	2	6
50	ему	252	56	2	4	3	I	5	4	3	6	II	4	2	7	I	4	I0	I2	2	7	2	6	4	7	I2	7	I	I0	2	3	4
51	теперь	237	5I	2	4	I	6	I	3	6	4	3	4	2	2	-	3	I2	4	I0	I	5	3	4	5	5	8	I2	3	2	I0	7
52	потом	234	54	3	4	3	I	5	2	2	8	2	2	4	I	2	2	II	7	I8	4	3	5	6	5	I0	8	I	4	4	3	5
53	ли	23I	5I	I	-	7	3	3	II	3	6	4	4	I	I	9	4	4	5	3	-	5	-	2	I4	3	6	6	-	4	3	7
54	него	2I8	53	I	2	-	I	I	6	I	4	3	3	I	4	6	3	9	8	8	2	5	5	5	7	6	II	4	5	2	I	3
55	даже	2I5	50	I	3	I	2	-	2	2	5	6	I	2	7	-	2	5	7	I3	3	3	5	7	I2	I0	4	6	I	-	7	2
56	ничего	2I5	52	I	2	2	-	I	2	I	I	3	5	I	3	-	2	I0	4	II	3	2	4	4	I6	9	7	4	6	2	I0	7
57	есть	2II	50	-	2	6	2	I	2	5	8	3	3	2	3	I	I	I5	I	7	3	2	-	7	I2	3	6	8	I	I	7	3
58	все	207	53	3	I	2	2	-	3	I	-	3	4	-	4	-	4	3	6	6	3	4	5	7	II	6	6	3	3	I	5	3
59	раз	207	53	2	2	3	4	I	4	4	2	-	4	I	4	5	5	4	3	5	2	5	I0	4	II	I0	7	-	3	3	I	2
60	их	205	48	2	6	4	3	-	2	-	2	2	4	I	I	-	6	8	4	I0	I	4	4	2	6	7	3	II	2	I	3	6
6I	до	I96	5I	I	I	-	2	2	2	2	2	2	2	2	2	3	-	II	5	I5	I	-	I	I	7	I2	3	6	5	-	2	6
62	ни	I92	47	-	-	I	5	-	-	5	2	6	4	5	2	-	3	5	3	I2	2	2	3	4	6	I3	5	I	5	5	I0	5
63	тебя	I90	45	-	I	I	4	-	-	9	I	-	8	2	-	3	I	II	4	3	6	7	4	-	8	7	8	-	I	5	I	-
64	тебе	I89	47	-	2	4	2	3	I	I0	2	2	5	2	I	2	I	6	I	-	I	4	-	-	I0	20	8	-	2	4	3	2
65	где	I87	50	I3	3	3	-	2	2	2	2	2	2	2	2	2	2	9	6	I	8	I	I	3	I	4	I2	7	-	2	4	5
66	один	I84	50	I	3	4	-	-	2	2	3	I	-	2	2	3	7	8	4	3	4	4	3	5	I3	II	9	3	4	2	2	2
67	пошёл	I73	53	2	-	4	-	-	4	3	2	3	-	2	I	3	I	5	2	I0	2	I	4	I	3	2	7	7	8	5	2	I
68	стал	I7I	53	2	2	I	2	2	2	5	3	I	-	3	4	2	2	4	8	II	-	5	I	3	5	2	4	8	8	4	2	I
69	если	I70	44	-	2	-	I	2	4	2	2	I	2	3	3	4	5	6	I	I	5	4	3	4	8	I	7	8	3	-	3	I

Ранг	Словоформа	№ текстов																									
		30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57
36	чего	1	9	5	3	1	8	9	11	12	13	7	3	3	–	3	5	10	5	3	4	3	6	10	11	3	11	10	3
37	они	4	19	2	9	5	1	3	–	6	11	4	6	3	3	4	3	3	6	16	18	5	1	3	6	4	10	9	–
38	нет	6	7	3	6	4	6	5	7	7	6	9	5	2	–	2	2	8	2	4	5	3	4	1	18	1	9	2	13
39	от	9	11	3	16	2	2	1	–	12	2	6	2	1	6	3	2	4	5	12	2	6	3	–	6	8	12	9	6
40	только	8	4	4	9	5	1	2	3	11	4	8	1	1	4	2	2	6	5	15	9	3	4	7	4	5	3	7	7
41	её	9	2	3	12	5	1	5	2	4	5	7	3	–	1	–	–	4	2	18	6	8	1	6	2	6	6	13	5
42	мне	6	4	–	8	3	1	3	10	12	4	1	4	–	9	–	2	4	6	5	6	–	2	6	6	7	6	1	3
43	опять	7	3	3	2	6	7	2	2	5	1	5	3	1	5	–	2	1	11	8	11	1	2	3	7	4	10	6	7
44	меня	5	5	2	6	2	2	1	7	10	4	2	1	1	5	1	3	5	5	2	3	2	4	6	13	3	5	4	3
45	когда	3	5	8	6	2	–	1	3	3	7	8	–	3	5	2	1	7	5	14	4	2	2	6	2	7	4	15	4
46	вы	2	2	–	6	1	4	1	2	5	3	2	2	1	–	2	–	2	13	7	1	1	1	–	15	5	12	5	17
47	был	8	3	12	3	–	2	3	1	6	–	2	3	5	2	3	2	5	20	5	5	–	3	5	1	–	2	5	1
48	уж	3	4	5	3	3	1	4	4	11	4	3	6	2	1	–	2	10	2	1	6	3	10	6	2	6	2	3	9
49	мы	3	3	7	1	2	–	2	2	1	–	2	6	9	12	7	1	9	3	5	9	1	2	1	5	6	2	–	6
50	ему	9	4	4	8	4	1	2	2	5	5	1	3	–	3	3	–	3	7	8	4	2	4	4	6	2	3	5	4
51	теперь	8	7	3	3	–	2	5	2	9	2	7	3	4	–	1	2	8	3	4	1	13	1	1	–	4	6	3	2
52	потом	4	3	8	6	4	1	2	1	1	2	6	3	1	3	2	1	3	1	7	13	1	–	5	3	5	3	5	6
53	ли	4	3	3	3	–	6	6	1	6	3	4	6	1	1	2	7	7	3	7	6	4	1	4	7	–	4	4	9
54	него	5	5	2	2	2	–	1	2	4	8	4	4	3	3	3	2	1	8	8	3	2	–	–	4	2	7	4	11
55	даже	1	1	2	6	–	6	–	1	2	6	5	7	1	1	2	7	4	7	6	2	–	–	1	7	4	1	11	5
56	ничего	5	3	4	2	2	3	2	2	3	3	6	2	–	2	–	–	3	8	3	1	2	1	3	7	5	4	8	8
57	есть	2	5	6	14	1	–	3	2	3	2	3	5	1	–	3	–	14	1	4	7	–	2	2	2	–	4	3	7
58	все	5	4	3	5	2	4	3	2	3	6	4	2	3	2	5	1	4	2	10	5	9	4	2	4	4	5	2	3
59	раз	4	4	7	6	3	–	2	6	4	3	3	1	3	4	–	1	4	1	6	4	2	6	2	3	8	3	1	5
60	их	1	12	3	2	1	–	8	2	–	15	–	3	2	4	–	–	6	4	6	4	5	1	1	4	–	9	5	2
61	до	4	6	3	7	2	3	–	–	2	2	5	1	1	2	1	2	3	1	1	12	2	5	1	1	12	6	6	4
62	ни	–	1	2	1	2	3	1	2	3	2	4	6	–	7	–	–	6	6	10	3	4	–	1	2	1	4	3	5
63	тебя	5	3	4	4	3	1	5	7	4	1	7	–	–	–	–	2	1	1	4	6	5	2	7	3	5	7	7	1
64	тебе	7	7	1	4	5	–	2	6	6	7	1	–	–	–	2	6	4	3	2	8	4	2	4	–	4	5	3	2
65	где	5	8	5	6	1	–	5	–	6	4	3	2	1	1	1	–	9	6	3	7	1	2	–	3	3	4	6	–
66	один	5	1	4	3	–	–	1	4	6	6	1	–	4	–	4	1	5	4	3	2	2	2	1	4	3	4	3	4
67	пошёл	2	3	6	6	2	2	2	5	5	8	3	2	2	1	1	1	1	5	3	3	5	2	3	3	1	6	2	3
68	стал	3	3	1	7	–	1	1	–	4	5	2	3	1	2	1	1	2	2	1	5	5	1	1	2	8	6	3	3
69	если	7	1	2	9	–	–	1	–	2	4	–	2	–	–	1	–	17	6	6	4	–	–	3	–	6	4	5	4

Ранг	Словоформ	Абсол. част.	Политекст.	№ текстов																										
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
70	кто	170	53	4	1	5	4	4	4	4	6	1	1	3	1	4	1	11	2	1	4	-	4	4	3	5	1	4	-	2	1	2
71	все	170	47	-	11	-	1	-	4	4	1	3	4	4	1	2	1	2	2	7	3	5	1	5	15	6	3	4	-	1	1	8
72	дело	169	51	-	2	5	1	2	3	1	4	3	3	-	2	1	-	3	3	11	1	1	5	2	6	10	2	4	3	1	6	1
73	будет	166	49	1	3	-	-	1	3	-	2	1	1	3	-	1	-	14	3	9	4	-	3	3	6	4	5	9	-	3	3	1
74	хорошо	166	47	2	1	-	1	1	1	1	2	-	5	5	4	1	1	7	2	11	2	6	7	3	9	-	5	3	-	1	3	7
75	нас	162	47	6	1	2	6	-	-	1	1	3	3	2	1	-	-	5	4	1	4	-	4	4	2	19	1	7	1	1	1	6
76	сам	162	50	2	-	2	4	1	1	3	3	2	2	1	3	-	2	4	-	1	5	1	3	2	11	5	4	4	7	1	2	5
77	мать	157	30	-	1	1	-	3	-	-	-	-	2	-	1	-	-	49	4	1	1	-	3	-	2	1	2	1	4	-	-	-
78	спросил	156	45	1	2	1	-	-	1	2	12	2	-	2	2	5	-	6	5	1	-	1	4	6	8	4	5	4	4	2	-	1
79	ведь	154	47	-	1	1	-	-	4	8	1	1	5	5	3	-	1	3	5	5	5	7	3	2	6	16	6	3	-	2	2	1
80	о	146	43	2	-	-	-	-	7	-	5	4	1	1	12	-	2	4	2	5	-	1	2	7	2	2	4	7	2	-	1	10
81	смотрел	142	45	1	6	-	3	3	-	-	1	4	-	1	6	1	-	1	-	4	4	4	10	2	3	2	3	1	8	3	2	-
82	можно	138	46	-	-	-	1	-	4	1	-	1	-	-	1	1	3	5	1	6	2	1	7	11	1	1	6	4	4	1	9	3
83	уже	136	47	-	-	1	-	-	2	1	2	-	1	1	2	2	1	6	6	5	4	4	1	-	8	3	10	4	5	6	3	2
84	себя	135	51	-	1	-	1	1	2	5	3	1	3	3	9	1	-	-	1	8	1	5	2	5	2	6	4	6	-	1	1	2
85	больше	134	44	-	3	1	-	1	3	5	2	4	1	4	1	-	1	4	2	6	2	3	-	1	7	11	2	6	2	-	3	4
86	под	132	49	1	3	1	1	3	3	4	2	2	-	1	-	1	2	4	1	11	2	3	3	-	1	-	5	7	3	4	1	1
87	хоть	130	46	2	3	1	5	4	3	1	1	2	2	4	1	3	2	8	2	1	3	1	1	2	2	4	9	-	3	-	6	1
88	люди	130	44	-	4	-	2	1	2	-	-	2	2	2	4	3	1	10	4	3	7	2	2	1	2	2	3	1	-	-	4	2
89	вас	129	39	4	2	2	-	-	-	2	3	4	9	1	2	2	-	4	-	-	2	3	3	2	3	-	5	5	-	-	4	2
90	конечно	128	42	2	1	-	-	-	1	-	3	-	3	5	5	-	-	4	2	5	-	2	5	5	5	3	7	8	3	2	4	-
91	или	127	50	1	2	-	1	1	1	2	2	2	1	2	2	1	2	8	2	7	3	2	1	2	3	5	2	1	-	1	3	2
92	были	126	46	2	2	1	1	-	2	6	4	1	3	3	2	1	4	5	2	7	1	1	1	1	4	2	1	1	4	-	3	7
93	знаю	125	45	1	2	-	3	-	1	4	-	5	-	3	1	1	-	1	2	2	3	2	3	3	1	7	3	4	-	1	2	8
94	человек	125	44	1	2	-	4	-	1	-	2	2	2	4	5	3	2	2	2	3	1	8	-	-	2	2	5	5	-	3	1	1
95	долго	122	44	-	4	2	2	4	1	-	-	-	1	3	3	4	1	2	2	3	4	1	4	2	4	2	2	2	6	1	1	5
96	без	121	46	-	3	1	2	2	3	-	1	-	3	2	-	4	2	2	2	5	-	3	1	1	4	2	6	9	1	2	5	4
97	Иван	121	12	-	-	-	-	-	-	-	-	22	-	4	-	-	-	-	-	4	-	-	-	-	-	1	-	-	-	39	1	-
98	как-то	121	48	-	3	4	2	2	1	2	3	1	-	2	2	2	-	4	2	7	2	3	2	4	4	3	5	6	1	1	3	2
99	говорит	120	33	-	-	10	2	1	-	3	1	1	-	2	-	8	3	5	-	3	12	-	-	-	5	8	6	1	2	-	2	4
100	сказала	119	36	-	1	9	3	1	1	-	9	1	2	1	-	-	-	2	2	6	-	-	-	2	5	-	4	2	2	-	2	11

Ранг	Слово-формы	№ текстов																											
		30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57
70	кто	3	2	–	7	3	5	I	4	2	I	I	–	2	I	5	2	3	I	6	9	4	4	3	5	I	3	3	2
71	про	I	I	5	5	I	I	2	4	2	I	7	–	–	–	–	2	IO	2	3	8	–	4	2	4	–	I	3	2
72	дело	3	5	3	IO	2	I	3	I	3	2	I	–	2	–	3	–	4	2	5	3	4	3	I	8	I	8	2	3
73	будет	3	3	–	4	I	3	2	4	8	5	5	I	2	I	2	2	I	2	–	I2	I	4	I	I	5	5	I	3
74	хорошо	IO	–	I5	5	–	I	–	I	2	4	I	3	I	–	2	I	3	I	I2	I	I	I	2	–	–	2	2	4
75	нвс	5	4	2	5	I	–	2	I	–	–	I	I	4	I2	I	–	IO	2	2	6	I	7	I	4	2	I	I	–
76	сам	2	5	4	IO	3	2	5	4	3	4	4	–	3	I	3	–	4	5	I	2	3	–	–	3	3	3	2	2
77	мать	9	8	I	–	–	I	I	–	20	–	I	I	–	I	–	–	I	–	5	6	I	24	–	–	–	–	–	I
78	спросил	5	3	2	2	–	I	2	2	I	6	–	5	I	–	–	I	–	2	I	3	–	3	I	6	I	IO	6	IO
79	ведь	2	3	2	4	I	3	2	2	4	3	4	4	–	–	–	–	2	2	I	I	2	I	2	2	I	–	2	IO
80	о	5	2	I	I	I	2	–	2	–	2	4	5	–	2	I	–	2	5	6	3	–	2	2	3	I	–	IO	I
81	смотрел	7	3	3	I	–	I	–	2	2	–	2	2	–	7	2	2	–	4	2	I	3	2	3	4	I	3	5	7
82	можно	2	I	2	3	I	7	I	–	5	4	–	2	I	I	–	I	4	I	2	4	2	2	2	2	9	–	2	3
83	уже	I	I	3	2	I	3	–	I	–	I	2	3	–	5	I	I	3	I	8	4	I	–	I	3	2	2	2	4
84	себя	–	4	2	6	3	I	2	I	3	I	I	2	I	I	3	2	2	I	4	3	2	2	I	4	I	4	2	2
85	больше	–	–	4	I	–	–	I	I	3	4	4	–	I	I	3	I	4	–	4	5	–	–	I	7	2	4	2	2
86	под	–	2	4	7	2	I	3	2	5	–	2	I	2	–	I	4	I	I	5	I	5	I	–	I	I	6	2	2
87	хоть	I	2	I	2	–	–	2	–	4	4	–	2	–	I	2	–	I	2	4	6	7	–	5	–	2	–	I	4
88	люди	5	I	3	3	–	–	I	5	4	–	3	–	I	–	I	–	3	–	5	6	5	2	2	I	I	4	3	5
89	вас	5	–	–	4	–	–	2	I	4	–	I	2	–	–	I	2	3	7	3	6	–	I	–	3	3	4	4	9
90	конечно	I	I	2	–	I	I	3	3	I	3	2	–	–	2	–	I	3	–	I	3	3	I	2	6	9	3	–	I
91	или	2	I	2	8	2	2	2	I	I	3	3	–	–	2	2	I	5	2	5	3	–	–	I	–	4	4	6	3
92	была	2	–	4	4	I	–	I	–	–	2	6	I	–	2	–	–	I	I	6	2	2	I	5	I	I	5	5	I
93	знаю	I	2	I	4	4	2	5	4	9	I	–	2	–	–	2	2	2	I	4	4	I	I	–	–	2	4	–	4
94	человек	I	2	4	6	3	–	I	2	–	–	I	–	–	–	3	3	5	2	2	6	I	I	2	6	2	4	5	–
95	долго	4	6	5	–	3	I	2	I	6	I	6	–	–	I	–	–	I	2	–	5	I	I	–	2	–	5	3	–
96	без	4	2	2	–	2	I	2	I	–	3	4	I	I	–	I	4	2	–	I	4	I	2	–	4	–	5	I	2
97	Иван	–	–	–	–	–	–	–	7	38	–	–	–	–	–	2	–	–	–	I	–	–	I	–	–	–	–	I	–
98	как-то	2	2	3	I	–	2	I	–	I	I	7	2	2	2	–	2	5	–	2	–	–	I	2	3	I	2	2	2
99	говорит	2	–	3	I	6	–	–	–	I	–	–	–	–	–	I	–	6	3	2	6	–	5	I	–	3	–	I	–
100	сказала	I	2	I	2	–	–	–	–	–	3	I	I	–	–	2	–	–	7	5	–	I	4	7	–	6	2	6	2

A FREQUENCY-DISTRIBUTED LIST OF SOME STORIES BY V.M. SHUKSHIN

Arkadi I. Kruchenkov

S u m m a r y

The frequency-distributed list is based on the word-forms taken from 57 stories by V. Shukshin. The list is given with the indication of the rank, the frequency of the wordforms and the volumes of the texts. The distribution of frequency in the texts is also taken into account. The complete volume of the studied texts includes 117,729 words. The given table shows 100 most frequency wordforms.

ОБ ОДНОМ СПОСОБЕ ДИФФЕРЕНЦИАЦИИ ТИПОВ НАУЧНО-ТЕХНИЧЕСКОГО ТЕКСТА

Н.С.Манасян

В последние годы неоднократно высказывалось мнение о необходимости применения дифференциальных признаков при описании стилевых подразделений языка.

Попытки выделить дифференциальные признаки осуществлялись разными способами. В качестве стилеразличительных параметров можно рассматривать разнообразные характеристики. Такие статистические исследования проведены многими стилистами на материале разных языков.

В качестве стилеразличительных признаков может выступать длина предложения в словоупотреблениях, как, например, в работах (Лесскис, 1963; Долежел, 1964). В этих и в ряде других работ сопоставление проведено по первичным данным (см., например, работы Г.Хердана, а также (Вольская, 1966; Головин, 1967; Микерина, 1967).

Исследование С.И.Кауфмана предполагает сопоставление не только по первичным признакам - здесь вводится коэффициент стиля,"характеризующий стиль в ограниченной области употребления в нем сказуемых и определений" (Кауфман, 1961, с. 5).

Работа (Шайкевич, 1968) несколько отстоит от перечисленных выше применением теории графов и использованием вспомогательного языка, который назван "средним языком".

В.С.Горевая (1974) же устанавливает систему корреляционных показателей, выступающих в качестве дифференциальных признаков стилевых рубрик. Стилевые рубрики при этом выделяются по принципу цели коммуникации в текстах на английском языке середины XX века. Автором принимается, что основным формальным методом анализа стилистических рубрик можно считать оппозиции. Основанием для последних считается общность структурного типа, противопоставление заключается в частотности употребления (или в наличии-отсутствии) данного структурного типа в разных стилевых рубриках.

В данной работе вопрос о дифференциальных признаках рассматривается на основе дифференцирующих возможностей таблиц сопряженности.

Исходным материалом служат данные табл.1, в которой

содержатся частоты встречаемости различных частей речи в разных типах научно-технического текста (ТНТТ).

Таблица I

Частотность употребления частеи речи в разных ТНТТ

Часть речи / ТНТТ	I	2	3	4	5	6	7	8
	Сущ.	Прил.	Арт.	Мест.	Глаг.	Нар.	Предл.	Союз.
I. Статья	239	76	88	26	I4I	24	92	50
2. Моногр.	290	9I	I74	23	I90	34	II7	7I
3. Учебник	292	I04	I38	5I	I86	I0	I28	45
4. Задачн.	35I	88	I52	22	238	I3	I30	57
5. Аннотац.	284	84	86	I6	I04	I7	90	28
6. Справ.	3II	86	I28	44	II4	47	I44	57
7. Реклам.	373	97	52	86	I42	39	96	64
8. Патент	345	II8	I22	29	I04	37	I4I	47
9. Деловое письмо	353	70	85	II0	I73	24	I45	69
I0. Бюллетень	368	93	79	55	I62	28	I28	42
II. Обсуж. докл.	279	80	I06	7I	I48	34	II7	37
И т о г о	3485	987	I205	533	I702	307	I328	567

П р и м е ч а н и е. Пренебрегаем статистикой некоторых частей речи, т.к. их частотность слишком мала, чтобы можно было расчитывать на содержательные выводы.

В соответствии с общепринятой терминологией (Кендалл, Стюарт, I973, с.745), эта таблица представляет собой $I \times J$ таблицу сопряженности для двух категоризованных признаков - части речи и ТНТТ, где I = II - число строк градаций признака "ТНТТ", J = I3 - число столбцов градаций признака "часть речи".

Выбирая различными способами группы столбцов из табл.

I, получаем ряд таблиц с I = II и разными J. Для каждой из них проверяем гипотезу независимости указанных признаков. С этой целью используется статистика

$$\chi^2 = n\left(\sum_{ij}\frac{x_{ij}}{x_{i*}x_{*j}} - 1\right), \qquad (1)$$

где i,j - номера строки и столбца соответственно, x_{ij} - число, стоящее в клетке таблицы на пересечении i-ой строки и j-того столбца, x_{i*} - сумма всех чисел в i-ой строке, x_{*j} - сумма всех чисел в j-ом столбце, n - сумма всех чисел в таблице.

Если выполняется гипотеза о независимости, то распределение величины χ^2 стремится при $n \to \infty$ к распределению с $(I-1)(J-1)$ степенями свободы (Кендалл, Стюарт, 1973, с.746). Критическая область для проверки независимости имеет вид

$$\chi^2 > \chi^2_{крит.}, \qquad (2)$$

где пороговое значение χ^2 определяется по таблицам распределения при заданном числе степеней свободы

$$\nu = (I-1)(J-1)$$

и уровне значимости α. В настоящей работе принято α = 0,05. В случае выполнения неравенства 2, т.е. если гипотеза о независимости отвергается, возникает вопрос об измерении силы связи между признаками. Поскольку в данной работе она предназначена для сопоставления таблиц с различными J, целесообразно в качестве меры связи использовать предложенную Крамером величину

$$C = \sqrt{\frac{\chi^2}{n\min(I-1)(J-1)}}, \qquad (3)$$

которая независимо от размера таблицы имеет достижимые нижнюю и верхнюю границы 0 и I.

Результаты вычислений приведены в табл.2. Данные в таблице расположены по убыванию величины C.

Автор пользуется случаем, чтобы выразить самую глубокую благодарность канд. физ.-мат. наук И.М.Сливняку за реализацию программы вычислений на ЭВМ и за помощь в выборе математической стратегии работы.

Таблица 2

№ пп	Вариант	ν	$\chi^2_{крит.}$	$\chi^2_{выч.}$	C
I.	Глагол, наречие	I0	I8	8I	0.20I
2.	Существ., мест.	I0	I8	I28	0.I78
3.	Существ., арт.	I0	I8	II7	0.I58
4.	Глагол, предлог	I0	I8	54	0.I33
5.	Существ., прилаг., арт., мест.	30	44	303	0.I28
6.	Существ., глагол	I0	I8	77	0.I22
7.	Предлог, союз	I0	I8	26	0.II7
8.	Все обследованные части речи	70	9I	464	0.076
9.	Существ., нар.	I0	I8	23	0.069
I0.	Существ., прилаг.	I0	I8	I9	0.065

Примечание. В графе "Вариант" указаны части речи, включенные в соответствующий вариант.

Как видим, для всех рассмотренных вариантов при уровне значимости $\alpha = 0{,}05$ гипотеза о независимости отвергается. Это означает, что признак "часть речи" является дифференцирующим, "типоразличающим" признаком для научно-технического текста.

Л И Т Е Р А Т У Р А

Вольская И.С. Дифференциальные признаки официально-делового стиля речи на синтаксическом уровне. - М., 1966.

Голодин Б.Н. О возможностях количественной характеристики речевых стилей. - Киев, 1961.

Горевая Н.С. Статистическое описание функционально-стилевых подразделений современного английского языка. - Калинин, 1974.

Доležel Л. Вероятностный подход к теории художественного стиля. ВЯ, 1964, № 2.

Кауфман С.И. Об именном характере технического стиля. ВЯ, 1961, № 5.

Кендалл М.Дж., Стюарт А. Статистические выводы и связи. - М., 1973.

Лесскис Г.А. О зависимости между размером предложения и характером текста. ВЯ, 1963, № 3.

Микерина Г.А. Некоторые статистические приемы лексико-морфологического описания функционального стиля. - Л., 1967.

Шайкевич А.Я. Опыт статистического выделения функциональных стилей. ВЯ, 1968, № 1.

ON A METHOD OF TECHNOLOGICAL TEXT DIFFERENTIATION

Narinay Manasyan

S u m m a r y

In the paper the problem of text differentiation is regarded on the basis of differential capabilities of correlation tables.

The part of speech occurence was calculated for 11 types of technological text (1000 words each). The numerical results are presented in a correlation table. Calculation of Chi-square between the two symptoms was realised on a computer. The results showed that part of speech symptom can differentiate different types of technological text.

О СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИКАХ РАНГОВЫХ РАСПРЕДЕЛЕНИЙ

Г.Я.Мартыненко

I. Введение.

Важной задачей квантитативной лингвистики является разыскание числовых характеристик, лаконично отображающих совокупные свойства лексических единиц, представленных в виде частотного словаря. В роли таких характеристик выступают: показатель ципфовского распределения, квантили этого распределения, его мода, энтропия, различные варианты степенных средних, индекс знакотип-знакоупотребление и некоторые другие (Горькова В.И., 1968; Яблонский А.И., 1975; Хайтун С.Д., 1983; Тулдава Ю.А., 1980). К этим характеристикам, как и к любым статистическим оценкам, предъявляются стандартные требования несмещенности, состоятельности и эффективности, вытекающие из предельных теорем теории вероятностей. Однако на практике многие статистические оценки таким требованиям не удовлетворяют. Это неприятное обстоятельство обычно объясняют негауссовостью распределений, изучаемых в гуманитарных дисциплинах (Яблонский А.И., 1975; Хайтун С.Д., 1983; Мартыненко Г.Я., Чарская Т.К.,1987). Внешним проявлением негауссовости является аномально большая вариация переменных и плохая сходимость экспериментальных величин к теоретическим при увеличении объема выборки.

2. О смысле рангового среднего.

Для изучения количественных закономерностей в тексте и корпусе текстов часто осуществляется многопредметное наблюдение, т.е. измеряются свойства разноименных объектов (например, лексических единиц), образующих относительно замкнутую систему (например, текст). Данные наблюдения при этом могут быть упорядочены несколькими способами.* Некоторые из них показаны в табл.I, из которой видно, что при переходе от кумулятивных рядов к ранжированным варианты и статистические веса меняются местами: варианты становятся значениями зависимой переменной

---

* О различных формах задания статистических распределений и связях между ними см.в работах (Хастингс К.,Пикок Дж.,1980; Мартыненко Г.Я.,1982; Нешитой В.В.,1986; Тулдава Ю.А.,1986).

Таблица I

Формы упорядочивания статистических данных

Формы упорядочивания	Статистические данные	
	Варианты	Статистические веса
Вариационный ряд (спектровое распределение)	Частота лексической единицы	Число лексических единиц с данной частотой
Кумулятивный ряд	Частота лексической единицы	Число лексических единиц с частотой, не превышающей данную
Декумулятивный ряд*	Частота лексической единицы	Число лексических единиц с частотой, превышающей данную
Возрастающий ранжированный ряд	Число лексических единиц с частотой, не превышающей данную (ранг "текущей" лексической единицы)	Частота лексической единицы
Убывающий ранжированный ряд	Число лексических единиц с частотой, превышающей данную (ранг "текущей" лексической единицы)	Частота лексической единицы

(т.е. функции), а статистические веса - значениями независимой переменной (т.е. аргумента). Обращает на себя внимание и то обстоятельство, что при переходе к ранжированным рядам накопляемые численности объектов "превращаются" в последовательность чисел натурального ряда, т.е. в ранговую последовательность .

* Термин "декумулятивный ряд" используется в эконометрии при построении распределений, в которых значения случайной величины образуют размер дохода, а статистическими весами являются численности лиц с доходом, превышающим данный (Ланге О., 1964)

Если прибегнуть к механическим параллелям, то ранговое распределение можно интерпретировать как некоторую, равную единице массу, распределенную по оси абсцисс (оси рангов) так, что в отдельных точках-координатах r_i сосредоточены соответствующие массы-вероятности p_i . Если механическую аналогию продолжить, что среднее, при вычислении которого, в качестве значений случайной переменной используются ранги, а в качестве статистических весов - частоты или их аналоги, может рассматриваться как центр тяжести рангового распределения.

Ранговое среднее обладает важной особенностью. Оно может рассматриваться не только как мера центральной тенденции но и как показатель концентрации-рассеяния "активности" единиц в совокупностях самой разнообразной природы: от неорганической до знаково-информационной. При этом минимальная концентрация (максимальное рассеяние) характерно для равномерного рангового распределения, в котором все элементы имеют одинаковую активность (например, частоту). Чем больше перепад между "головой" и "хвостом" рангового распределения, тем выше уровень концентрации относительно равномерного распределения. Иначе говоря, это распределение может рассматриваться как эталон, относительно которого измеряется концентрация.

3. Некоторые математические свойства ранговых статистик.

Пусть мы имеем возрастающий ранжированный ряд, задаваемый таблицей:

r_i	f_i
1	f_1
2	f_2
...	...
$r-1$	f_{r-1}
r	f_r

Ранговое среднее этого ряда равно:

$$\bar{r}_e = \frac{1 \cdot f_1 + 2 \cdot f_2 + \ldots + (r-1) \cdot f_{r-1} + r\, f_r}{\sum f_i} \qquad (I)$$

Преобразуем возрастающий ранжированный ряд в убывающий:

r_i	f_i
1	f_r
2	f_{r-1}
...	...
$r-1$	f_2
r	f_1

Ранговое среднее этого ряда равно;

$$\bar{r}_y = \frac{1 \cdot f_r + 2 \cdot f_{r-1} + \ldots + (r-1) f_2 + r f_1}{\Sigma f_i} \quad (2)$$

Просуммируем $\bar{r}_y$ и $\bar{r}_e$:

$$\bar{r}_e + \bar{r}_y = \frac{(r+1) f_1 + (r+1) f_2 + \ldots + (r+1) f_{r-1} + (r+1) f_r}{\Sigma f_i} =$$

$$= \frac{(r+1) \Sigma f_i}{\Sigma f_i} = r+1$$

Итак, сумма средних, вычисленных для убывающего и возрастающего ранжированных рядов, равна численности разноименных единиц совокупности, увеличенной на единицу(или максимальному рангу плюс единица).*

Перестроим теперь оба ранжированных ряда в кумулятивный и декумулятивный ряды:

Кумулятивный ряд:

f_i	r_i
f_1	1
f_2	2
...	...
f_{r-1}	$r-1$
f_r	r

Декумулятивный ряд:

f_i	r_i
f_1	r
f_2	$r-1$
...	...
f_{r-1}	2
f_r	1

* На первый взгляд это свойство ранговых средних может показаться тривиальным. Действительно, если объект находится в"очереди" на K-ом месте от начала,то он находится на $[(n+1)-k]$-ом месте от хвоста (n - длина очереди). Это относится к любому индивидууму, стоящему в очереди. Но ранговое среднее характеризует не отдельный индивидуум, а совокупные свойства очереди. Поэтому установленное свойство не представляется столь уж очевидным.

Определим средние для обоих рядов:

$$\bar{f}_к = \frac{f_1 \cdot 1 + f_2 \cdot 2 + \ldots + f_{z-1} \cdot (z-1) + f_z \cdot z}{\Sigma r_i} \quad (3)$$

$$\bar{f}_у = \frac{f_1 \cdot z + f_2 \cdot (z-1) + \ldots + f_{z-1} \cdot 2 + f_z \cdot 1}{\Sigma r_i} \quad (4)$$

Сумма $\bar{f}_к$ и $\bar{f}_у$ равна:

$$\bar{f}_к + \bar{f}_у = \frac{(z+1)\Sigma f_i}{\Sigma r_i} = \frac{(z+1)\Sigma f_i}{z \cdot \frac{z+1}{2}}$$

После несложных преобразований получаем:

$$\bar{f} = \frac{\bar{f}_у + \bar{f}_к}{2}$$

где $\bar{f}$ – среднее арифметическое частот.

Итак, средняя частота равна полусумме средних, вычисленных для кумулятивного и декумулятивного рядов.

Покажем теперь, что между средними, вычисленными для рядов различного типа существует функциональная связь.

Разделим выражение (3) на выражение (I):

$$\frac{\bar{f}_к}{\bar{r}_в} = \frac{\Sigma f_i r_i}{\Sigma r_i} \cdot \frac{\Sigma f_i}{\Sigma r_i f_i} = \frac{\Sigma f_i}{\Sigma r_i} = \frac{\Sigma f_i}{z \cdot \frac{z+1}{2}} = \frac{2}{z+1} \cdot \bar{f}$$

Итак: $\frac{\bar{f}_к}{\bar{r}_в} = \frac{2}{z+1} \cdot \bar{f}$

Соответственно: $\frac{\bar{f}_у}{\bar{r}_у} = \frac{2}{z+1} \cdot \bar{f}$

Следовательно: $\frac{\bar{f}_у}{\bar{r}_у} = \frac{\bar{f}_к}{\bar{r}_в} = \frac{2}{z+1} \cdot \bar{f}$

Остановимся теперь на одном свойстве ранговых дисперсий.

Чтобы не усложнять вывод, рассмотрим ранжированный ряд, состоящий из трех элементов.

Для возрастающего и убывающего рядов имеем соответственно:

$$\sigma^2_{r_в} = \frac{\sum r_i^2 \cdot f_i^в}{\sum f_i} - \left(\bar{r}_в\right)^2$$

$$\sigma^2_{r_y} = \frac{\sum r_i^2 f_i^y}{\sum f_i} - \left(\bar{r}_y\right)^2$$

Выясним, что представляет собой разность двух дисперсий:

$$\sigma^2_{r_в} - \sigma^2_{r_y} = \left[\frac{\sum r_i^2 f_i^в}{\sum f_i} - \frac{\sum r_i^2 f_i^y}{\sum f_i}\right] + \left[\left(\bar{r}_y\right)^2 - \left(\bar{r}_в\right)^2\right]$$

Обозначим первое слагаемое через y, второе - через z.

$$y = \frac{1^2 \cdot f_1 + 2^2 f_2 + 3^2 \cdot f_3}{f_1 + f_2 + f_3} - \frac{1^2 f_3 + 2^2 f_2 + 3^2 f_1}{f_1 + f_2 + f_3} = \frac{8f_3 - 8f_1}{f_1 + f_2 + f_3},$$

$$z = \left(\bar{r}_y\right)^2 - \left(\bar{r}_в\right)^2 = \left(\bar{r}_y - \bar{r}_в\right)\left(\bar{r}_y + \bar{r}_в\right) = \frac{(4f_1 + 4f_2 + 4f_3)(2f_1 - 2f_3)}{f_1 + f_2 + f_3}$$

$$= \frac{8f_1 - 8f_3}{f_1 + f_2 + f_3}$$

Из того, что $y + z = 0$, следует, что дисперсии равны. Распространив этот вывод на ранжированные ряды любой протяженности, приходим к заключению, что дисперсия ранжированного ряда в отличие от рангового среднего не зависит от его ориентации.

4. Ранговые параметры некоторых теоретических распределений.

Рассмотрим теперь ранговые параметры нескольких распределений, широко используемых в статистической практике: распределение Ципфа-Парето, распределение Вейбулла и один из вариантов логистической функции. Различные формы задания этих распределений приведены в табл.2.

Начнем с закона Ципфа-Парето, имеющего для убывающей формы рангового распределения следующий вид:

$$Z(r) = \frac{(K+1)^{\frac{1}{\alpha}}}{r^{\frac{1}{\alpha}}}$$

Этой формуле (см.табл.2) в форме, описывающей вариационный ряд (спектровое распределение), соответствует формула

Таблица 2

Формы аналитического задания некоторых теоретических распределений

	Распределение Ципфа–Парето	Распределение Вейбулла	Логистическая функция
I. Кумулятивный ряд	$(K+1)\left(1-\frac{1}{x^{\gamma}}\right)$	$(K+1)\left(1-e^{-cx^{\gamma}}\right)$	$\frac{(K+1)x^{\gamma}}{x^{\gamma}+q^{\gamma}}$
2. Вариационный ряд	$(K+1)\gamma x^{-(\gamma+1)}$	$(K+1)c\gamma x^{\gamma-1}e^{-cx^{\gamma}}$	$\frac{(K+1)q^{\gamma}x^{\gamma-1}}{\frac{1}{\gamma}(x^{\gamma}+q^{\gamma})^{2}}$
3. Возрастающий ранжированный ряд	$(K+1)^{\frac{1}{\gamma}}\left[(K+1)-z\right]^{-\frac{1}{\gamma}}$	$\left[\frac{1}{c}\ln\frac{K+1}{(K+1)-z}\right]^{\frac{1}{\gamma}}$	$\left(\frac{K+1}{(K+1)-z}-1\right)^{\frac{1}{\gamma}}q$
4. Декумулятивный ряд	$\frac{K+1}{x^{\gamma}}$	$(K+1)e^{-cx^{\gamma}}$	$\frac{(K+1)q^{\gamma}}{x^{\gamma}+q^{\gamma}}$
5. Убывающий ранкированный ряд	$(K+1)^{\frac{1}{\gamma}}z^{-\frac{1}{\gamma}}$	$\left[\frac{1}{c}\ln\frac{K+1}{z}\right]^{\frac{1}{\gamma}}$	$\left(\frac{K+1}{z}-1\right)^{\frac{1}{\gamma}}q$

$$f(x) = \frac{(K+1)\gamma}{x^{\gamma+1}}$$

которая может быть получена из классической формулировки закона Ципфа-Парето (Хайтун С.Д., 1983; Липкин М.И., 1972):

$$f(x) = \frac{\gamma}{x_0}\left(\frac{x_0}{x}\right)^{\gamma+1}$$

где x_0 - наименьшее теоретическое значение случайной переменной. При $x_0 = 1$ (а именно такое значение x_0 фигурирует чаще всего в информатике, науковедении, лингвистике, биологии и др. дисциплинах) эта формула приобретает вид:

$$f(x) = \frac{\gamma}{x^{\gamma+1}}$$

Это выражение справедливо для плотности вероятности. Если же пользоваться абсолютными частотами, то в эту формулу нужно ввести дополнительный множитель ($K+1$). Единица в этом множителе появляются потому, что сумма двух ранговых средних (в убывающем и возрастающем ранжированном рядах), как было установлено выше, равна увеличенному на единицу объему совокупности.

Определим математическое ожидание и дисперсию для убывающей формы рангового распределения закона Ципфа-Парето:

$$M(z) = \frac{\int_0^{K+1} \frac{(K+1)^{\frac{1}{\gamma}}}{z^{\frac{1}{\gamma}}} \cdot z\,dz}{\int_0^{K+1} \frac{(K+1)^{\frac{1}{\gamma}}}{z^{\frac{1}{\gamma}}}\,dz} = \frac{(K+1)^2}{2-\frac{1}{\gamma}} : \frac{K+1}{1-\frac{1}{\gamma}} = \frac{(K+1)\left(1-\frac{1}{\gamma}\right)}{2-\frac{1}{\gamma}}$$

$$\sigma^2(z) = \frac{\int_0^{K+1} \frac{(K+1)^{\frac{1}{\gamma}}}{z^{\frac{1}{\gamma}}} \cdot z^2\,dz}{\int_0^{K+1} \frac{(K+1)^{\frac{1}{\gamma}}}{z^{\frac{1}{\gamma}}}\,dz} - \left[M(z)\right]^2 =$$

$$= \frac{(k+1)^2\left(1-\frac{1}{\gamma}\right)}{3-\frac{1}{\gamma}} - \frac{(k+1)^2\left(1-\frac{1}{\gamma}\right)^2}{\left(2-\frac{1}{\gamma}\right)^2} = \frac{(k+1)^2\left(1-\frac{1}{\gamma}\right)}{\left(3-\frac{1}{\gamma}\right)\left(2-\frac{1}{\gamma}\right)^2}$$

Определим также среднее квадратическое отклонение и коэффициент вариации:

$$\sigma(r) = \frac{(k+1)\sqrt{1-\frac{1}{\gamma}}}{\sqrt{3-\frac{1}{\gamma}}\left(2-\frac{1}{\gamma}\right)}$$

$$v(r) = \frac{1}{\sqrt{\left(1-\frac{1}{\gamma}\right)\left(3-\frac{1}{\gamma}\right)}}$$

Математическое ожидание и дисперсия закона Ципфа существует при $\gamma > 1$.

Аналогичным способом могут быть определены и параметры двух других распределений. Итоговые результаты сведены в таблицу 3.

Остановимся теперь на способе построения индексов концентрации ранговых распределений.

В формулах ранговых средних

$$\bar{r}_y = \frac{r+1}{2} \cdot \frac{\bar{f}_g}{\bar{f}}$$

$$\bar{r}_e = \frac{r+1}{2} \cdot \frac{\bar{f}_k}{\bar{f}}$$

$\frac{r+1}{2}$ есть величина постоянная, равная ранговому среднему равномерного распределения. Поэтому, для того, чтобы получить обобщенную меру концентрации или рассеяния можно одну из этих формул разделить на эталонную величину $\frac{r+1}{2}$, но то же самое можно сделать, разделив $\bar{r}_y$ на $\bar{r}_e$:

$$R = \frac{\bar{r}_y}{\bar{r}_e} = \frac{\bar{f}_g}{\bar{f}_k} \leq 1 .$$

R всегда меньше единицы, т.к. $0 < \bar{f}_g < \bar{f}$, $\bar{f} < f_k < 2\bar{f}$.

Таблица 3

Ранговые параметры некоторых теоретических распределений

	Функция Ципфа–Парето	Функция Вейбулла	Логистическая функция
Математическое ожидание $M(z)$	$\frac{(K+1)(1-\frac{1}{\gamma})}{2-\frac{1}{\gamma}}$	$\frac{K+1}{2^{\frac{1}{\gamma}+1}}$	$\frac{K+1}{2}\left(1-\frac{1}{\gamma}\right)$
Дисперсия $\sigma^2(z)$	$\frac{(K+1)^2(1-\frac{1}{\gamma})}{(3-\frac{1}{\gamma})(2-\frac{1}{\gamma})^2}$	$(K+1)^2 \times \times\left(\frac{1}{3^{\frac{1}{\gamma}+1}}-\frac{1}{2^{2(\frac{1}{\gamma}+1)}}\right)$	$\frac{K+1}{12}\left[1-\left(\frac{1}{\gamma}\right)^2\right]$
Стандартное отклонение $\sigma(z)$	$\frac{(K+1)\sqrt{1-\frac{1}{\gamma}}}{\sqrt{3-\frac{1}{\gamma}}\,(2-\frac{1}{\gamma})}$	$(K+1)\times \times\sqrt{\frac{1}{3^{\frac{1}{\gamma}+1}}-\frac{1}{2^{2(\frac{1}{\gamma}+1)}}}$	$\frac{K+1}{2\sqrt{3}}\sqrt{1-\left(\frac{1}{\gamma}\right)^2}$
Коэффициент вариации $v(z)$	$\frac{1}{\sqrt{(1-\frac{1}{\gamma})(3-\frac{1}{\gamma})}}$	$\sqrt{\frac{2^{2(\frac{1}{\gamma}+1)}}{3^{\frac{1}{\gamma}+1}}-1}$	$\frac{1}{\sqrt{3}}\sqrt{\frac{1-\gamma}{1+\gamma}}$
Индекс концентрации H	$\frac{1}{\gamma}$	$\frac{2^{\frac{1}{\gamma}+1}-2}{2^{\frac{1}{\gamma}+1}-1}$	$\frac{2}{\gamma}$

Чем ближе R к единице, тем равномернее распределение частоты между единицами совокупности, т.е. в данном случае концентрация минимальная. Желательно, однако, иметь прямопропорциональную зависимость. Чтобы достичь желаемого нужно из единицы вычесть R.

Итак:

$$H = 1-R = 1-\frac{\overline{z}_g}{\overline{z}_a} = 1-\frac{\bar{f}_g}{\bar{f}_K},$$

где H – индекс концентрации.

Чем ближе H к единице, тем большая масса частот концентрируется в голове убывающего рангового распределения.

На основе данных, помещенных в табл.3, можно сделать следующие выводы:

При $\gamma \to \infty$ параметры всех теортических распределений устремляются к параметрам равномерного распределения.

$$M(r) \to \frac{K+1}{2}\ ,\quad \sigma^2(r) \to \frac{(K+1)^2}{12}\ ,\quad v(r) \to \frac{1}{\sqrt{3}}\ ,\quad H \to 0$$

При $\gamma = 1$ функция Вейбулла превращается в показательную с математическим ожиданием $M(r) = \frac{K+1}{4}$, в два раза меньшим математического ожидания закона равномерной плотности. Следовательно, показательное распределение тоже может рассматриваться как эталонное.

Обращает на себя внимание то, что индекс концентрации непосредственно входит в уравнение закона Ципфа.

Это совершенно однозначно (не метафорически) раскрывает содержательный смысл коэффициента γ или его обратной величины. Этот коэффициент может быть получен непосредственно из данных наблюдения и при условии ципфовости распределения может непосредственно служить мерой его концентрации.

Обсудим теперь характер зависимости рангового среднего от объема выборки. Сделаем это на материале частотных словарей. Попутно обсудим проблему зависимости объема словаря от объема текста.

5. Зависимость рангового среднего от объема выборки.

К сожалению, в публикациях составителей частотных словарей редко приводятся частотно-ранговые сетки, отражающие закономерности перестройки структуры словаря при возрастании объема текста или объема выборки. Удовлетворимся не слишком полными данными о динамике перестройки словаря, которые приводятся в работах П.М.Алексеева и Г.Кучеры (Алексеев П.М., 1965; Kučera H. , 1967). Данные П.М.Алексеева относятся к тематически однородному подъязыку (подъязык электроники), в то время как словарь Г.Кучеры построен на материале текстов, относящихся к самым разнообразным сферам человеческой деятельности: науке, технике, политике, религии, спорту, публицистике, праву, беллетристике и т.д. Иначе говоря, словарь Г.Кучеры построен с претензией на охват всех сфер функционирования английского языка.

Нами была выдвинута гипотеза, что при определенном объеме выборки величина рангового среднего стабилизируется, достигая определенного асимптотического уровня. Проверку этой гипотезы можно осуществить путем построения нескольких аппроксимирующих функций. Среди них выбирается та, которая наилуч-

шим образом согласуется с эмпирическими данными.

В качестве аппроксимирующих мы выбрали следующие функции асимптотического роста:

степенную – $y = K - Kx^{-c}$,

экспоненциальную – $y = K - Ke^{-cx}$,

логарифмическую – $y = K - \frac{K}{\ln cx}$,

комбинацию степенной с экспоненциальной (функция Вейбулла) – $y = K - Ke^{-cx^d}$,

комбинацию степенной с логарифмической – $y = K - \frac{K}{(\ln cx)^d}$.

Данные наблюдения выравнивались по перечисленным функциям с помощью метода наименьших квадратов. Применение этого метода значительно облегчается тем, что каждая из этих функций путем однократного или повторного логарифмирования может быть преобразована в линейную. Линейные варианты перечисленных функций имеют вид:

$$\ln \frac{K}{K-y} = c \ln x,$$

$$\ln \frac{K}{K-y} = cx,$$

$$\frac{K}{K-y} = c \ln x,$$

$$\ln\ln \frac{K}{K-y} = \ln c + d \ln x,$$

$$\ln \frac{K}{K-y} = \ln c + d \ln\ln x.$$

Система нормальных уравнений линейной зависимости очень проста и ее решение относительно неизвестных параметров не вызывает никаких затруднений. Однако в нашем случае один из параметров (асимптота K) входит в состав зависимой переменной, а это исключает использование метода наименьших квадратов в его чистом виде. Нами был найден следующий выход из сложив-

шейся ситуации. Асимптоте мы задавали конкретные значения через определенный шаг, начиная с максимальных значений ранговой средней и объема словаря, достигнутых в эмпирическом ряду. Далее, на каждом шаге при фиксированной величине асимптоты с помощью метода наименьших квадратов вычислялись численные значения других параметров, и на каждом же шаге фиксировалась средняя величина квадратов отклонений эмпирических данных от теоретических. После этого среди этих величин выбиралась минимальная (для каждой аппроксимирующей функции). Среди полученных пяти минимальных величин в свою очередь выбиралась минимальная. Тот закон, для которого был получен этот минимум, и считался наилучшим приближением к эмпирическому распределению. Применительно к нашим опытным данным такой минимум был получен для функции Вейбулла, причем как для рангового среднего так и для объема словаря. На рис.I на материале частотного словаря подъязыка электроники приведены результаты работы этой оптимизационной процедуры.

Теоретические и экспериментальные значения рангового среднего и объема словаря, включая прогностические данные, приведены в табл.4, а соответствующие графики на рис.2 и 3.

Ранговые средние тематически и функционально ограниченных подъязыков стабилизируются при сравнительно небольшом объеме выборки. Это хорошо видно ра рис.2. Убедимся в этом, рассчитав объем выборки при заданном размере рангового среднего. Пусть последний составляет 99% от максимально возможного. Тогда, осуществив переход от функции Вейбулла к обратной функции, получаем:

$$\mathcal{L} = \left(\frac{1}{c}\, e_n \frac{N_{max}}{N_{max} - N}\right)^{\frac{1}{d}}.$$

Подставив в эту формулу соответствующие данные, получаем: $\mathcal{L}$ = I90,9 тыс.словоупотреблений. Это означает, что уже при выборке, равной I90,9 тыс.словоупотреблений, ранговое среднее составляет 99% от максимально возможной.*

---

* Полученные нами данные могут рассматриваться как дополнительный фактор, оправдывающий объем в 200 тыс.словоупотреблений, принятый за эталон при составлении частотных словарей в Общесоюзной группе "Статистика речи" (Алексеев П.М., I975).

Таблица 4

Экспериментальные и теоретические характеристики частотных словарей

Объем выборки (тыс. словоупотреблений)	Объем словаря		Ранговая средняя	
	эксперимент.	теоретич.	эксперимент.	теоретич.
Частотный словарь П. М. Алексеева				
50	5399	542I	62I	626
I00	7853	7827	75I	732
I50	936I	94I0	754	763
200	I0582	I0565	772	777
500(прогноз)	-	I3672	-	780
I000(прогноз)	-	I4880	-	780
Частотный словарь Г. Кучеры				
I0,0	3009	3026	546	539
I0I,0	I3706	I3530	I256	I332
253,5	23655	23676	I937	I846
I0I4,2	50406	50597	28I0	2833
I0000 (прогноз)	-	I05070	-	I4447
I00000 (прогноз)	-	II2500	-	74940

Пусть теперь достигнутый уровень рангового среднего составляет 99,9% от максимально возможного. В этом случае объем выборки должен составить 32I,0 тыс.словоупотреблений.

Стабилизация ранговых средних тематически и функционально неоднородной выборки наступает значительно позднее: 99% - ный уровень рангового среднего достигается при $\mathcal{L}$ = 50,2 млн. словоупотреблений, а практически предельный 99,9% -ный уро - вень - при $\mathcal{L}$ =I28, I млн.словоупотреблений. Из этого следует, что для достижения предельной величины рангового среднего объем выборки в общелитературном языке должен быть как мини - мум на два порядка больше, чем в тематически ограниченном подъязыке.

Что касается максимального объема словаря, то он дости -

гается в частотном словаре подъязыка электроники значительно позднее, чем для рангового среднего: при $\mathcal{L}$ = 2,38 млн.словоупотреблений . Напомним, что для рангового среднего эта цифра значительно меньше (321 тыс.словоупотреблений). Обращает на себя внимание и тот факт, что в общелитературном языке весь словарь "вычерпывается" при $\mathcal{L}$ = 40,4 млн. словоупотреблений, т.е. значительно раньше, чем достигается предельный уровень рангового среднего. Это является дополнительным аргументом в пользу слабой сходимости ранговых статистик к постоянной величине в неоднородной совокупности.

6. Выводы.

I. При изучении информационных потоков в качестве статистик ранговых распределений естественно пользоваться счетными средними, при расчете которых в роли значений независимой переменной выступают ранги, а в роли статистических весов – частоты (или их аналоги), соответствующие этим рангам.

2. Ранговые средние являются показателями концентрации (или рассеяния) "активности" единиц, из которых образовано ранговое распределение. Равномерное ранговое распределение, обладающее минимальной концентрацией, может рассматриваться как эталон, относительно которого исчисляются концентрация и рассеяние.

3. Ранговые средние обнаруживают быструю сходимость к предельным величинам, но только тогда, когда исследуемая совокупность единиц качественно однородна. В разнородной совокупности (частотный словарь языка в целом) скорость сходимости очень мала.

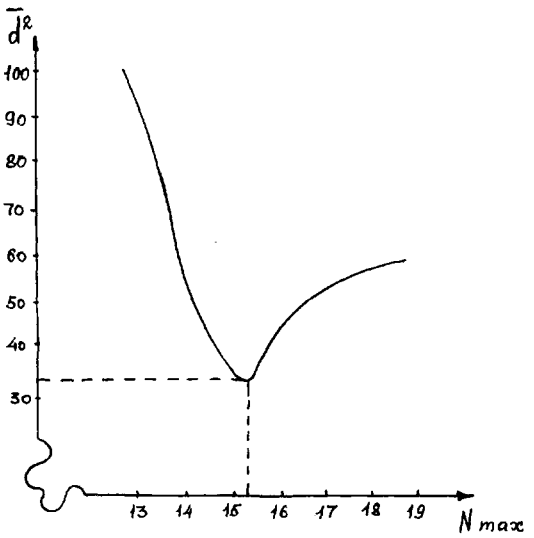

Рис.I. Величина среднего квадрата отклонений ($\bar{d}^2$) теоретических значений распределения Вейбулла от экспериментальных при различных значениях максимального объема словаря (N_{max}).

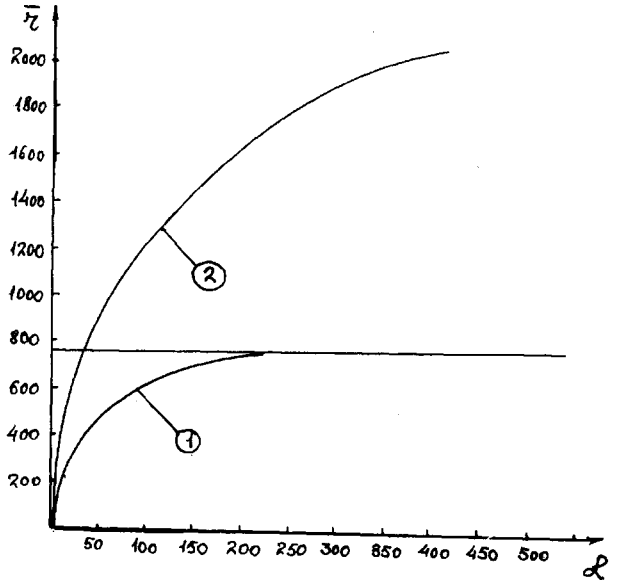

Рис.2. Зависимость рангового среднего ($\bar{r}$) от объема выборки (ℓ).

I – частотный словарь современного английского языка (составитель Г.Кучера):

2 – английский частотный словарь подъязыка электроники (составитель П.М.Алексеев):

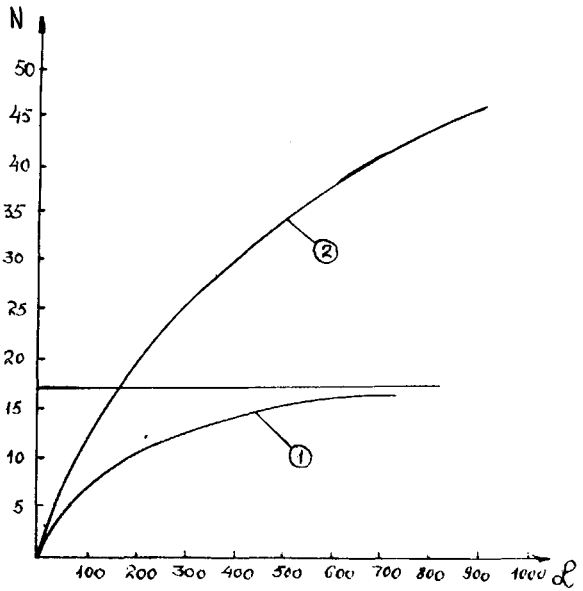

Рис.3. Зависимость объема словаря (N) от объема выборки (ℓ) .

I – частотный словарь современного английского языка (составитель Г.Кучера):

2 – английский частотный словарь подъязыка электроники (составитель П.М.Алексеев):

Л И Т Е Р А Т У Р А

Алексеев П.М. Частотный словарь английского подъязыка электроники: Автореф.дис...канд.филол.наук. – Л.,1965.

Алексеев П.М. Статистическая лексикография (типология, составление и применение частотных словарей). – Л.: ЛГПИ им. А.И.Герцена, 1975.

Горькова В.И. Ранговые распределения на множествах научно-технической информации.-Научно-техническая информация. Сер.2, 1968, № 5, с.5-11.

Ланге О. Введение в эконометрику. М.: Прогресс, 1964.

Липкин М.И. Кривые распределения в экономических исследованиях. – М.: Статистика, 1972.

Мартыненко Г.Я. Типология лингвостатистических распределений. – Учен.зап.Тартуского ун-та. Вып.628. Лингвостатистика и вычислительная лингвистика. Тарту, 1982, с.103-118.

Мартыненко Г.Я., Чарская Т.К. К вопросу о гауссовости и негауссовости лингвостатистических распределений. – В кн.: Структурная и прикладная лингвистика. Вып.3. – Л.,1987, с.63-76.

Нешитой В.В. О взаимосвязи ранговых распределений со спектровыми.- Научно-техническая информация. Сер.2, 1968, 10, с.19-24.

Тулдава Ю.А. К вопросу об аналитическом выражении связи между объемом словаря и объемом текста. – Учен.зап.Тартуского ун-та, Вып. 549. Лингвостатистика и квантитативные закономерности текста. Тарту, 1980, с.113-144.

Тулдава Ю.А. О частотном спектре лексики текста. – Учен.зап. Тартуского ун-та. Вып.745. Квантитативная лингвистика и автоматический анализ текста. Тарту, 1986, с.139-162.

Хайтун С.Д. Наукометрия. Состояние и перспективы. – М.: Наука, 1983.

Хастингс Н., Пикок Дж. Справочник по статистическим распределениям. Пер. с англ. – М.: Статистика, 1980.

Яблонский А.И. Стохастические модели научной деятельности. В кн.: Системные исследования. Ежегодник. М., 1975, с.5-42.

Kučera H., Francis W.N. Computational analysis of present-day American English. – Providence, 1967.

RANK DISTRIBUTION STATISTIC CHARACTERISTICS

Grigory Martynenko

S u m m a r y

It is proposed to use rank means as statistics of empiric rank distributions. To calculate the former ones random variable is substituted for ranks and statistic weights for frequencies (or their analogs) corresponding to these ranks. Rank means are characteristics of non-uniformity (concentration and scattering) of lexical units in different zones of frequency dictionary, uniform rank distributions being an etalon(a standard) according to which concentration and scattering are calculated. Rank distributions show rapid convergence to limiting values with the increase of sample volume but not until the population being investigated is homogeneous, e.g. it covers the dictionary of subject limited sublanguage.

СТАТИСТИКА ТЕРМИНОЛОГИЧЕСКИХ СЛОВОСОЧЕТАНИЙ В АНГЛИЙСКОМ ПОДЪЯЗЫКЕ ЯДЕРНОЙ ФИЗИКИ

Е.В.Мурмуридис

В квантитативной лингвистике научного текста (КЛНТ), исследующей реализуемые в этом тексте системные лингвистические объекты, важная роль принадлежит изучению терминологических систем подъязыков, к которым относятся конкретные тексты. Большинство работ в области статистической лексикографии ограничивается однословной терминологией хотя бы в силу относительно простого в формальном плане выделения слова в тексте. Идентификация терминологических словосочетаний сложнее и по формальным, и по содержательным причинам, требует лучшего знания предметной области и более частого обращения к информанту-специалисту. Кроме того, на доступных для индивидуального исследования объемах текстовой выборки статистика слов естественным образом дает бóльшие значения частот, чем статистика словосочетаний при одинаковой длине текста. Поэтому в первом случае и сами данные выглядят достовернее, и их количество, отвечающее заданному частотному порогу, значительно превышает корпус данных для второго случая.

Применение ЭВМ в статистике словосочетаний вообще и терминосочетаний в частности сдерживается отсутствием возможности задавать формальные границы сочетаний в тексте. Правда, способы машинной регистрации сочетаний требуемой длины с опорой на известное слово известны давно. Предлагаются процедуры для машинного опознания в тексте и терминологических сочетаний.[X] Тем не менее, в КЛНТ пока еще многое приходится связывать с домашинной ручной обработкой, разметкой текста, а трудоемкому немашинному способу получения частотных словарей терминов, особенно терминосочетаний, пока еще нет приемлемой альтернативы.

[X] Пиотровский Р.Г., Билан В.Н., Боркун М.Н., Бобков А.К. Методы автоматического анализа и синтеза текста. - Минск: Вышэйшая школа, 1985. - 222 с.

Ниже приводится частотный список терминологических словосочетаний, составленный вручную на материале зарубежных англоязычных периодических изданий. Объем выборочного корпуса равен 100 тыс. словоупотреблений в 100 текстах по 1 тыс. словоупотреблений в каждом. К терминологическим словосочетаниям отнесены не только термины в узком смысле, но и некоторые сочетания слов, характерные для научного текста.

По разделам ядерной физики тексты распределяются следующим образом:

Теоретические проблемы. Физика нейтрино - 5, общая классификация элементарных частиц - 10, взаимодействие ядерных частиц - 17, теория строения атомного ядра - 17.

Практические проблемы. Частицы низких энергий - 14, частицы высоких энергий - 17, нейтронная физика - 6, ядерные реакции (естественные и искусственные) - 12.

Техника. Термоядерная энергетика - 4, ядерные реакторы (техника и технология) - 2, совмещенные реакторы (синтез-деления)- 2.

В выборке зарегистрированы 5829 разных словосочетаний, из которых в список помещены 955 сочетаний с частотами не менее 3. Не вошли в приводимый список 943 сочетания с частотой 2 и 3931 сочетание с частотой 1. Слева от сочетания указан его частотный ранг, а справа - частота.

Одно из свойств терминологического сочетания состоит в том, что с уменьшением его частоты все более конкретизируется передаваемое им понятие, а описываемое научное явление характеризуется все более детально. Это легко видеть на примере терминов, содержащих общий элемент cross section (90)[x]; differential cross section (25); inclusive cross section (15); fusion cross section (10); experimental cross section (8); calculated cross section (6); hadronic cross section, production cross section, reduced cross section (4); cali-

---

[x] Здесь и далее в тексте цифры указывают на частоту в обследованной выборке.

bration cross section, reaction cross section, resulting cross section (3). Многие сочетания передают неделимые в данной предметной области понятия; они фиксируются в специальном словаре и функционируют как цельные образования. Если принять за основу цельность научного понятия в пределах данного терминополя, то элементы таких"синтагматических цепочек" связаны прежде всего парадигматическими отношениями. Целые группы сочетаний не воспринимаются каждое как нечто целое, передающее одно понятие. Словари обычно приводят звенья таких сочетаний, и вопрос о границах этих звеньев может быть дискуссионным. Это можно видеть на примере, таких, скажем, сочетаний, как renormalization group improved ladder model(3), или (nnm) MeV proton elastic scattering data (3). Здесь синтагматические отношения в целом сочетании, как кажется, преобладают над парадигматическими, хотя четкую границу между первыми и вторыми проследить трудно.

Ранговое распределение терминосочетаний [x]

i	F	m	i	F	m	i	F	m
I	93	I	I5-I6	2I	2	72-9I	I0	I0
2	90	I	I7-I8	20	2	92-I0I	9	I0
3	42	I	I9-20	I9	2	I02-I32	8	3I
4	38	I	2I	I8	I	I33-I65	7	33
5	36	I	22-24	I7	3	I66-235	6	68
6	35	I	25-29	I6	5	236-323	5	88
7	32	I	30-4I	I5	I2	324-507	4	284
8-I0	25	3	42-47	I4	6	508-955	3	448
II	24	I	48-56	I3	9	956-I898	2	943
I2	23	I	57-62	I2	6	I899-5829	I	393I
I3-I4	22	2	63-7I	II	9			

[x] i - ранг, F - частота, m - количество сочетаний с частотой F .

1	in the case	93
2	cross section	90
3	in terms of	42
4	to carry out	38
5	of the order of	36
6	for example	35
7	final state	32
8-10	in agreement with, as a function of, differential cross section	25
11	angular distribution	24
12	energy dependence	23
13-14	in the range, detecting system	22
15-16	by a factor of, ground state	21
17-18	excitation energy, elastic scattering	20
19-20	experimental data,matrix element	19
21	decay mode	18
22-24	coupling constant, in detail, in ... form	17
25-29	in n (num) dimensions, form factor, excitation function, wave function, in particular	16
30-41	to account for, high energy, kinetic energy, low energy, generating functional (n), angular momentum, present paper, to point out, analyzing power, experimental result, inclusive cross section, mass singularity	15
42-47	transverse momentum, order of magnitude, branching ratio, dispertion relation, bound state, gauge supersymmetry	14
48-56	in good agreement, on the basis of, level density, structure function, fission product, dimentional regularization, total cross section, partial wave, present work	13
57-62	scintillation counter, volume integral, prototype model, decay scheme, gauge theory, momentum transfer	12
63-71	decay amplitude, jet axis, muon capture, strong interaction, target nucleus, imaginary potential, mass spectrum, absolute value, theoretical value	11
72-91	leading logarithmic (log) approximation, mass	10

forward amplitude,TDHF calculation, for the case, it is the case, for comparison, potential depth, detailed description, photon detector, momentum distribution, neutron multiplicity distribution, infrared divergence, higher order effect, delayed particle emission, binding energy, lower energy, separation energy, incident energy, equation of motion, evolution equation, neutron evaporation, fission event, auxiliary field, pion field, ultraviolet finiteness, fission fragment, heavy fragment, parton structure function, fission isomer, effective lagrangian, dashed line, energy loss, ETC model, average number, composite operator, leading order, to order (n), imaginary part, secondary particle, vector particle, relative phase, new physics, Argand plot, transition probability, energy range, for the reason, energy resolution, vortex ring, level scheme, calculated cross section, neutron shell, isotope shift, isomer shift, relative size, classical solution, phase space, magnetic spectrometer,isomeric state, magnetic tape, finite temperature, broken color gauge theory, Fierz-Pauli Theory, massless theory, charged track, natural uranium, average value, starting value, bifunctional variable

236-323 reasonable agreement, nucleon-nucleon amplitude, scattering amplitude, forward angle, Born approximation, impulse approximation, one-loop approximation, under the assumption, singlet NG boson, wire spark chamber, ionization chamber, electric charge, topological charge, standard choice, in coincidence with, shorter component, force constant, running coupling constant, rigid core, relativistic correction, uranium cost, Čerenkov counter, gamma-ray counter, trigger counter, selection criterion, neutral current, hadronic decay, covariant derivative, in the forward direction, spin distribution, quark-confinement effect, operator product expan-

sion, good fit, light fragment, in the framework, Green function, axial gauge, spherical harmonics, range counter hodoscope, WT identity, experimental information, of ... kind, conservation law, proton lifetime, in the high-energy limit, invariant mass, reaction machanism, compuond nucleus reaction mechanism, standard model, compound nucleus, heavier nucleus, a number of, exchange current operator, neutrino oscillation, muon pair, previous paper, optical (-) model parameter, Higgs particle, virtual photon, effective potential, pion-optical potential, Woods-Saxon potential, vector analyzing power, in ... practice, in principle, experimental procedure, (num) - step process, physical quantity, rms (r.m.s.) radius, in the energy range, wide range, gamma ray, induced reaction, angular resolution, data sample, electron scattering, inelastic scattering, multiple scattering, odd signature, first excited state, initial state, compound system, data acquisition system, hydrogen target, grand unified theory, under the transformation, quark field variable, fractional cumulative yield

324-507 complete analysis, first approxamation, muonic 4
atom, on the average, error bar, bremsstrahlung beam, incoming beam, Nambu-Goldstone boson, heavy vector boson, symmetry breaking, microscopic calculation, breeding capability, in the present case, spark chamber, decay channel, exclusive channel, quantum chromodynamics, D3D code, transmission coefficient, proton-proton collision, fuel composition, experimental conditions, under ... conditions, in conjunction with, in ... context, meson exchange contribution, exchange current contribution, spin-saturated core, radioactive correction, threshold Čerenkov counter, exchange current, scalar current, smooth curve, solid curve, yield curve, raw data, muon decay, proton decay, radial dependence, depending on, electron detector, proton-recoil detector,

Argand diagram, Feynman diagram, Δ-N mass difference, at ... distance, flux distribution, proton-recoil distribution, ultraviolet divergence, neutron EDM, associated electronics, Auger emission, c.m. energy, resonance energy, total energy, 'Hooft equation, non-relativistic estimate, given event, event by event, experimental evidence, large-N expansion, vector field, effective force, nucleon-nucleon force, functional form, of the form, normal frequency, LEU fuelling, nuclear wave function, (num)-particle correlation function, strength function, Feynman gauge, gauge group, charmed hadron, decay heat, backward hemisphere, liquid hydrogen, lepton interaction, residual interaction, gauge invariance, hadronic jet, implanted layer, low-lying level, in the chiral limit, scaling limit, in the soft limit, straight line, (num) cm long, analyzing magnet, bending magnet, group manifold, effective mass, heavy mass, present measurement, previous measurement, maximum likelihood method, magnetic model, nuclear model, parton model, present model, shell model, standard SU(2) U(1) model, theoretical model, outgoing momentum, prompt neutron, light nucleus, mass number, phase parameter, flavor part, beam particle, charmed particle, final-state particle, Coulomb phase, oriented plaquette, data point, saddle point, starting point, spin-orbit potential, hadron production, inclusive production, computer program, quark propagator, convergence property, bound proton, light quark, final-state quark, over a ... range, mass region, forward dispertion relation, numerical result, preliminary result, present result, theoretical result, experimental sample, quark-quark scattering, hadronic cross section, parton cross section, production cross section, reduced cross section, data set, right-hand side, energy signal, time-of-flight signal, neutron

energy spectrum, Υ- ray spectrum, isobaric analog state, nuclear state, field strength, resonance strength, strength of the resonance, Dirak string, neutron structure, detailed study, conformal supergravity, gauge symmetry, coordinate system, fissioning system, polarized target, contact term, mass term, pole term, relativistic term, (num) - body term, spontaneously broken gauge theory, supersymmetric theory, rise time, storage time, positive track, EXP transformation, Gamow-Teller transition, partial transition, fast MI transition, for... use, fixed value, likelihood value, minimum value, nominal value, optimized value, field variable, scaling variable, production vertex, higher-order partial wave, chain yield, fractional yield

508-955 pion absorbtion, geometrical accelerator, accounting for, close agreement, complex amplitude, elementary amplitude, exchange amplitude, overlap amplitude, photoproduction amplitude, UPE amplitude, data analysis, M-matrix analysis, OMP analysis, phenomenological analysis, elastic scattering analysis, phase shift analysis, polarization analyzer, azimuthal angle, c.m. angle, polar angle, experimental arrangement, off- and on-shell mass assignment, intermediate-range attraction, ion beam, pion beam, tagged photon beam, high energy behavior, negative bias, control board, charged boson, NG boson, Weinberg-Salam type Higgs boson, spontaneous breaking, explosive carbon burning, beam burst, detailed calculation, lowest-order calculation, present calculation, time-dependent Hartree-Fock calculation, candidate for reaction, radiative capture, massive case, relativistic case, two-quark case, multiwire chamber, thin-walled ionization chamber, electromagnetic charge, integer charge, external charge, magnetic charge, nuclear charge, computer code, finite-difference code, correlation 3

coefficient, triple coincidence, direct comparison, small computer, boundary condition, Dirac condition, under consideration, normalization constant, linear constraint, quadratic constraint, particle content, higher-order contribution, significant contribution, fan cooler, core cooling, impact coordinates, absorption correction, higher-order correction, two-particle correlation, proportional counter, in the course of, germanium crystal, gauge-variant U(1) current, axial vector current, weak charged current, equilibrium curve, fuel cycle, external fuel cycle, leading cycle, forward data, photoabsorption data, polarization data, recent data, elastic scattering data, (num) MeV proton elastic scattering data, systematic data, beta decay, suitably defined, by definition, power density, single-particle density, state density, trial density, Department of energy, strongly dependent, well depth, functional derivative, high-energy description, fission fragment detector, muon detector, neutron detector, non-iterative diagram, self-energy diagram, in the direction, large discrepancy, detailed discussion, helpful discussion, energy dissipation, charge distribution, energy distribution, experimental distribution, jet axis angular distribution, multiplicity distribution, phase spase distribution, IR divergence, electroweak doublet, technifermion doublet, weak SU (2) doublet, to draw a conclusion, reimannian dynamics, collective effect, field effect, proton pairing effect, three-particle effect, density matrix element, reduced matrix element, proton emission, beam energy, condensation energy, given energy, photon energy, nucleon energy, differential equation, ladder model Bethe-Salpeter equation, Schrödinger equation, tadpole equation, wave equation, human error, theoretical estimate, electron event, gamma ray event, registered event, numerical example, natural parity exchange, unnatural parity

exchange, nuclear excitation, perturbation expansion, perturbative expansion, earlier experiment, previous experiment, deep inelastic scattering experiment, triggered experiment, to the extent, Lorentzian extrapolation, of a factor of, vertex factor, light fermion, gauge field, massive field, scalar field, infrared finite, delayed fission, prompt fission, better fit, kinematical fit, Drell-Yan formula, energy fraction, c.m. frame, laboratory frame, reference frame, rest frame, in the ... frame, within the framework, rolling friction, coefficient function, correlation function, distribution function, calculated excitation function, inclusive correlation function, likelihood function, beta strength function, F_2 structure function, color threshold function, pion wave function, light-cone gauge, real gluon, particle data group, final-state hadron, pulse height, forward hemisphere, particle identification, dispersion integral, initial integral, electromagnetic interaction, inelastic interaction, neutrino interaction, weak interaction, BRS invariant (adj), chiral invariant (adj), individual jet, opposite jet, to our knowledge, QCD lagrangian, scattering length, hyperfine level, nuclear level, within the limits, external line, full line, spectrometer magnet, boson mass, Higgs boson mass, jet mass, light mass, density matrix, accurate measurement, cancelling mechanism, GIM mechanism, computer memory, core memory, finite difference method, beam type method, vacuum metric, vibrational mode, internal vibrational mode, zero mode, collective model, continuous model, fuel-cycle model, Gross-Neveu model, Higgs model, little bag model, Pati-Salam model, preequilibrium model, proximity friction model, (num) - quark model, quasirealistic model, renormalization group improved ladder model, statistical model, TC

model, calculated moment, four momentum, loop momentum, static monopole, charged multiplicity, overall charged multiplicity, in the nature, daughter nucleus, deformed nucleus, final nucleus, parent nucleus, effective number, light fermion number, neutron number, Heisenberg operator, first order, quark-antiquark pair, recent paper, asymmetry parameter, geometrical parameter, level (-) density parameter, polarization parameter, radius parameter, spin parity, atomic part, finite part, fast particle, scalar particle, scattered particle, short-lived particle, beam path, prompt peak, horizontal plane, muon polarization, recoil proton polarization, target polarization, complex conjugate pole, firing position, fitted position, average potential, linear potential, one-loop effective potential, bending power, model prediction, theoretical prediction, uranium price, a priori, decay probability, probability of decay, probability of fission, decay process, scattering process, data processing, hadronization product, reaction product, nuclear production, lepton pair production, production of pions,fitting program, indocrination program, videotape program, in progress, vertical quark propagator, decay property, electromagnetic property, fast proton, detector pulse, gate pulse, initial quark, off-shell quark, in question, platinum radiator, charge radius, interaction radius, at ... range, muon capture rate, experimental ratio, mixing ratio, cosmic ray, charged current neutrino reaction, scattering reaction, breeder reactor, geometrical reconstruction, angular region, uniform field region, closely related, coincidence relation, in representation, giant resonance, fine structure resonance, (num) keV resonance, detector responce, at rest, final result, similar result, typical result, control room, se-

lection rule, data taking run, event sample, selected sample, industrial scale, ordinary hadronic mass scale, at ... scale, calibration cross section, photoneutron cross section, predicted cross section, reaction cross section, resulting cross section, separability of variables, parameter set, setting up (ger), stable equilibrium shape, triaxial equilibrium shape, mass shift, phase shift, resonance frequency shift, in short, master signal, start signal, relatively small, LENZ solution, multimonopole solution, parameter space, proton-arm spectrometer, secondary particle spectrometer, antineutrino spectrum, particle spectrum, proton spectrum, virtual photon spectrum, integer spin, hyperfine splitting, containment spray, spread of (in) likelihood, at ... stage, to stand for, random start, in ... state, multiparticle state, residual state, Weinberg state, yrast state, isospin structure, jet structure, nuclear structure, nuclear study, systematic study, subensemble of events, nuclear surface, color symmetry, rheonomic symmetry, electron synchrotron, light system, low (-) energy tail, data tape, butanol target, carbon target, counter technique, missing-mass technique, at ... temperature, at zero temperature, Born term, constant term, quadratic term, optical theorem, free field theory, lattice theory, Yang-Mills theory, (num) cm thickness, of (num) mm thickness, secondary track, direct transfer, high momentum transfer, Fourier transform, (num) keV transition, radiationless transition, experimental uncertainty, t- and s-channel unitarity, in units of, best value, empirical value, final value, high value, initial value, maximum value, mean value, measured value, necessary Q-value, necessary Q_2 value, parameter value, previous value, vacuum value, kinematic variable, main vertex, primary vertex, CP violation, fiducial volume, sensitive

volume, plane wave, full width, resonance width, total width, total hadronic decay width, time window, cumulative yield, fission yield.

STATISTICS OF COMPOUND TERMS IN ENGLISH TEXTS ON NUCLEAR PHYSICS

ELENA MURMURIDIS

S u m m a r y

The problems of identifying and registering special terms, and compound ones in particular, are of essential significance for quantitative linguistics of scientific text. Different means of compiling word-combination frequency dictionaries (FDs) have been offered including computer-aided techniques. But "manual production" of a compound-terms FD still remains one of the most reliable ways to get adequate statistics of text terminological structure.

An FD of compound terms based on 100,000 running words is presented here covering frequencies not less than 3.

МАТЕМАТИЧЕСКИЕ МОДЕЛИ РОСТА СЛОВАРЯ И ИНФОРМАЦИОННЫХ ПОТОКОВ

В.В.Нешитой

При решении различных задач прогнозирования необходимо знать характер зависимости между изучаемыми величинами, например, между числом названий книг и брошюр или заявок на изобретения и временем; числом разных дескрипторов и числом заиндексированных документов; числом разных информационных запросов и числом абонентo-запросов (т.е. с учетом их повторяемости); числом разных слов частотного словаря и объемом выборки и т.д.

Целью настоящей статьи является отыскание уравнений, описывающих различного рода кривые роста из областей информатики и математической лингвистики и решение на их основе некоторых задач.

В качестве моделей кривых роста могут быть использованы различные формулы. Во-первых, это уравнения, найденные для описания кривых роста н о в ы х событий (под "н о в ы м" понимается любое из n разных событий, составляющих полную группу, при первом его появлении от начала испытаний). Во-вторых, это обобщенные плотности и функции распределения, задающие системы непрерывных распределений (при условии снятия ограничений, наложенных на параметры кривых распределения). Все эти модели содержат не более четырех параметров. Использование общих моделей значительно облегчает поиск выравнивающей кривой, хотя и не освобождает исследователя от глубокого анализа статистических данных и знания свойств указанных кривых. Как правило, частная модель содержит меньше параметров, чем их имеется в общей модели. Рассмотрим четыре основные системы кривых роста.

I. Система I кривых роста задается формулой (см.Нешитой В.В., I986):

$$y = \int_0^n \left(1 - e^{-xp(t)}\right) dt , \qquad (1)$$

где y - математическое ожидание числа разных событий (разных слов), наступающих при x испытаниях (появляющихся в выборке объемом x словоупотреблений); $p(t)$ - непрерывная плотность распределения, аппроксимирующая вероятности p_k (k=I,2,...,

n) разных событий, составляющих полную группу.

При заданной плотности $p(t)$ по формуле (I) может быть рассчитана кривая роста новых событий. Все такие кривые удовлетворяют условиям

$$\left.\begin{aligned} &\frac{dy}{dx} = 1 \quad при \quad x \to 0 \\ &\frac{dy}{dx} \to 0 \quad при \quad x \to \infty \end{aligned}\right\} \qquad (2)$$

В качестве плотности распределения $p(t)$ в формулу (I) может входить любая плотность, заданная на интервале $0 < t < n$, где число разных событий n может быть как конечным, так и бесконечным. Например, это могут быть обобщенные плотности

$$p(t) = N t^{\gamma-1}\left(1 - \alpha u t^{\beta}\right)^{\frac{1}{u}-1}, \qquad (3)$$

$$p(y) = \frac{N}{y}(\ln y)^{\gamma-1}\left(1 - \alpha u \ln^{\beta} y\right)^{\frac{1}{u}-1}. \qquad (4)$$

Метод нахождения оценок параметров α, β, γ, u обобщенных плотностей (3), (4) изложен в работе автора (Нешитой В.В., I985).

Недостатком этой системы кривых роста является то, что интеграл (I), как правило, не выражается конечным числом элементарных функций. Кроме того, необходимо знать закон распределения вероятностей разных событий, составляющих полную группу.

2. Система П (а, б) кривых роста.

Система Па кривых роста в общем виде задается уравнением

$$\frac{dy}{dx} = 1 - \bar{F}(y) = 1 - \bar{F}(x), \qquad (5)$$

где $\bar{F}(y), \bar{F}(x)$ – функции распределения вероятностей новых событий.

Пусть $\bar{F}(y), \bar{F}(x)$ задаются формулами (Нешитой В.В., I986)

$$\bar{F}(y) = 1 - (1 - \alpha u y)^{\frac{1}{u}}, \qquad (6)$$

$$\bar{F}(x) = 1 - \left[1 - \alpha(u-1)x\right]^{\frac{1}{u-1}}. \qquad (7)$$

Тогда система Па кривых роста на основании (5) и (6), (7) будет описываться уравнением

$$y = \frac{1}{\alpha u}\left[1 - (1-\alpha(u-1)x)^{\frac{u}{u-1}}\right], \qquad (8)$$

которое удовлетворяет условиям (2).

Кривые роста, заданные общим уравнением (8), разделяются на типы. К I типу относятся кривые при $0 < u < \infty$; ко II типу - при $u \to 0$; к III типу - при $-\infty < u < 0$.

Оценки параметров α, u кривых I типа рассчитываются по формулам

$$\alpha = \frac{1}{x^2}\sum_{m\geq 1} m^2 y_m - \frac{1}{x}, \quad u = \frac{1}{\alpha n},$$

где $x = \sum_{m\geq 1} m y_m$, $n = \sum_{m>0} y_m$, y_m - число разных событий, наступающих при x испытаниях ровно m раз; n - число разных событий, составляющих полную группу.

Кривая роста относится к I типу, если $y_{m=0} > 0$.

В случае пуассоновского процесса ($u \to$ I) уравнение (8) примет более простой вид

$$y = \frac{1}{\alpha}\left(1 - \frac{1}{e^{\alpha x}}\right).$$

При $u \to 0$ из (8) имеем кривую роста II типа

$$y = \frac{1}{\alpha}\ln(1+\alpha x),$$

параметр α которой находится по методу **итераций**

$$\alpha_{i+1} = \frac{1}{y}\ln(1+\alpha_i x).$$

В случае кривых роста III типа, а также I и II типов, оценки параметров α, u могут быть найдены упрощенным методом по трем известным из опыта величинам: $x, y, y_{m=1}$, т.е. по объему выборки x, количеству наступивших разных событий y и количеству событий с частотой $m = 1$. При этом тип кривой устанавливается с помощью формулы

$$\frac{x}{y} = \frac{\frac{x}{y_{m=1}} - 1}{\ln\frac{x}{y_{m=1}}},$$

справедливой для кривых II типа. Если эмпирическое отношение $(x/y)^* = x/y$, то кривая роста относится ко II типу.

При $(x/y)^* < x/y$ - к I типу. При $(x/y)^* > x/y$ - к III типу.

Система IIб кривых роста получается из системы IIа введением замены $x = \ln X$, $y = \ln Y$. Тогда

$$\ln Y = \frac{1}{\alpha u}\left[1-\left(1-\alpha(u-1)\ln X\right)^{\frac{u}{u-1}}\right] \qquad (9)$$

или в дифференциальной форме

$$\frac{dY}{dX} = \frac{Y}{X}\left(1-\alpha u \ln Y\right)^{\frac{1}{u}} = \frac{Y}{X}\left[1-\alpha(u-1)\ln X\right]^{\frac{1}{u-1}} \qquad (10)$$

Дифференциальное уравнение (10) при различных значениях параметра u дает ряд формул для описания кривых роста новых событий (например, новых слов в тексте), однако удобными для практического использования являются те из них, которые позволяют в явном виде выражать параметр α через переменные X, Y. Это формулы при $u = I/2$, $u = -I$ (см. табл.I).

Чтобы иметь более широкий набор кривых роста (с равным интервалом $\Delta u = 0{,}75$), на основании формул с параметрами $u = I/2$, $u = -I$ были получены еще две формулы, которые соответствуют значениям $u \approx I{,}25$, $u \approx -0{,}25$. Они тоже приведены в табл.I.

Проверка кривых роста новых слов в текстах показала, что параметр α не является постоянной величиной, однако эмпирические точки в системе координат $(\ln X;\ \alpha)$ ложатся на прямую

$$\alpha = \alpha_0 + \kappa \ln X \ . \qquad (11)$$

Параметры α_0, κ являются параметрами текста и легко находятся по графику.

Эти же формулы оказались пригодными и для описания кривой роста новых слов в случайной выборке, но параметры выборки не совпадают с параметрами текста $(\alpha_{0в} < \alpha_{0т}\ ,\ \kappa_в > \kappa_т)$. Примером такой кривой является кривая роста числа разных дескрипторов $Y = f(X)$, где $X = Dh$ - число использованных дескрипторов при индексировании D документов; h - среднее число дескрипторов в одном документе.

Система Пб кривых роста новых слов в выборке Таблица I

Параметр u	Уравнение кривой роста	$\alpha = \alpha_0 + \kappa \ln X$	Оценки параметров выборки α_0, κ
$u \approx 1{,}25$	$Y = X^{1/X^{\alpha/2}}$	$\alpha = \frac{2}{\ln X} \ln \frac{\ln X}{\ln Y}$	$\kappa = \frac{2}{\ln^2 X}\left(1 - \ln \frac{\ln X}{\ln Y} - \frac{Y_{m=1}}{Y}\frac{\ln X}{\ln Y}\right)$ $\alpha_0 = \alpha - \kappa \ln X$
$u = 0{,}5$	$Y = X^{1/(1+\frac{\alpha}{2}\ln X)}$	$\alpha = \frac{2}{\ln X}\left(\frac{\ln X}{\ln Y} - 1\right)$	$\kappa = \frac{2}{\ln^2 Y}\left[\left(\frac{\ln Y}{\ln X}\right)^2 - \frac{Y_{m=1}}{Y}\right]$ $\alpha_0 = \alpha - \kappa \ln X$
$u \approx -0{,}25$	$Y = X^{1/\sqrt{1+\alpha \ln X}}$	$\alpha = \frac{1}{\ln X}\left[\left(\frac{\ln X}{\ln Y}\right)^2 - 1\right]$	$\alpha_0 = \frac{2}{\ln X}\left[\frac{Y_{m=1}}{Y}\left(\frac{\ln X}{\ln Y}\right)^3 - 1\right]$ $\kappa = \frac{1}{\ln^2 Y} - \frac{1}{\ln^2 X} - \frac{\alpha_0}{\ln X}$
$u = -1$	$Y = e^{\frac{1}{\alpha}\left(\sqrt{1+2\alpha \ln X} - 1\right)}$	$\alpha = \frac{2}{\ln Y}\left(\frac{\ln X}{\ln Y} - 1\right)$	$\kappa = \frac{2}{\ln^2 Y}\left[1 - \frac{Y_{m=1}}{Y}\left(2\frac{\ln X}{\ln Y} - 1\right)\right]$ $\alpha_0 = \alpha - \kappa \ln X$

Оценки параметров выборки могут быть найдены по трем величинам: $X, Y, Y_{m=1}$. При этом используется формула В.М.Калинина (1964, с. 247)

$$\frac{d^m y}{dx^m} = (-1)^{m+1} \frac{m!}{x^m} y_m ,$$

которая при $m = 1$ дает

$$\frac{dy}{dx} = \frac{y_{m=1}}{x} . \qquad (12)$$

На основании (12) и соответствующих уравнений кривой роста с учетом равенства (11) были найдены оценки параметров выборки, которые приведены в табл. 1.

Из приведенных в табл.1 формул лишь две последние ($u \approx -0{,}25$, $u = -1$) удовлетворяют условиям (2), при этом от параметра k зависит наибольший объем словаря. Кривая роста с параметром $u \approx -0{,}25$ при $X \to \infty$ дает: $Y_{max} = e^{\sqrt{1/k}}$. Для кривой с параметром $u = -1$ величина $Y_{max} = e^{\sqrt{2/k}}$. При $k \to 0$ в обоих случаях $Y_{max} \to \infty$. Следовательно, параметр k может служить показателем лексического разнообразия (богатства) текста. При равных значениях k бóльшим лексическим разнообразием обладает кривая с меньшим значением параметра α_0.

Первые две формулы ($u \approx 1{,}25$, $u = 0{,}5$) описывают кривые, которые вначале возрастают, затем убывают (при весьма больших значениях X). Следовательно, их можно использовать в качестве моделей кривых роста новых слов при ограниченных значениях X ($X < 10^6 \div 10^7$).

Формулы системы IIб кривых роста содержат всего два параметра, оценки которых легко находятся графическим методом либо рассчитываются по трем величинам $X, Y, Y_{m=1}$, при этом не требуется знание закона распределения разных слов по частоте их употребления в текстах.

3. Система III кривых роста задается формулами

$$y = NF(t), \qquad (13)$$

$$y = N[1 - F(t)]. \qquad (14)$$

где N - некоторый параметр, причем, $N = y_{max}$. Свойства кривых роста этой системы полностью определяются свойствами функции распределения $F(t)$.

Формула (I3) может описывать, например, количество разных статей по определенной теме, опубликованных в первых t журналах, при условии, что последние упорядочены по убыванию количества таких статей, а также количество заболеваний при эпидемиях за время t от начала эпидемии.

4. Система IУ (а, б, в) кривых роста

Система IУ кривых роста строится на основе обобщенных распределений. Вводя другие обозначения переменных и освобождаясь от ограничений, накладываемых на кривые распределения, можем записать следующие уравнения для описания различного рода кривых роста

$$\text{IV}\,a: \quad y = N x^{\gamma-1}(1-\alpha u x^{\beta})^{\frac{1}{u}-1}; \tag{I5}$$

$$\text{IV}\,б: \quad y = N e^{\gamma x}(1-\alpha u e^{\beta x})^{\frac{1}{u}-1}; \tag{I6}$$

$$\text{IV}\,в: \quad \ln y = N(\ln x)^{\gamma-1}(1-\alpha u \ln^{\beta} x)^{\frac{1}{u}-1}. \tag{I7}$$

Система IУ кривых роста является наиболее широкой системой, включающей кривые самой разнообразной формы. Она включает также некоторые кривые, принадлежащие другим системам.

4.I. Система IУа кривых роста новых слов

Проверка показала, что общая формула (I5) может достаточно точно описывать кривые роста новых слов при $\gamma = 2$, $N = I$:

$$y = x(1-\alpha u x^{\beta})^{\frac{1}{u}-1}. \tag{I8}$$

Наиболее подходящей оказалась формула при $u = -I$ $(0 < \beta < I/2)$

$$y = \frac{x}{(1+\alpha x^{\beta})^2}, \tag{I9}$$

которая применима при $(I0^3 \div I0^4) < x < (I0^7 \div I0^8)$ словоупотреблений. Оценки параметров текста находятся из уравнения прямой

$$\ln\left(\sqrt{\frac{x}{y}} - 1\right) = \ln\alpha + \beta \ln x. \tag{20}$$

Оценки параметров выборки можно найти по трем величинам: $x, y, y_{m=1}$

$$\beta = \frac{1 - \frac{y_{m=1}}{y}}{2\left(1 - \sqrt{\frac{y}{x}}\right)} = \frac{\sqrt{\frac{x}{y}}\left(1 - \frac{y_{m=1}}{y}\right)}{2\left(\sqrt{\frac{x}{y}} - 1\right)}, \tag{21}$$

$$\alpha = \frac{1}{x^{\beta}}\left(\sqrt{\frac{x}{y}} - 1\right). \tag{22}$$

4.2. Система IУв кривых роста новых слов

Уравнение (I7) хорошо описывает кривую роста новых слов при $\gamma = 2$, $N = I$, $(-I \le u < 0)$, т.е. система IУв кривых роста задается общей формулой

$$\ln Y = \ln X \left(1 - \alpha u \ln^{\beta} X\right)^{\frac{1}{u} - 1}. \tag{23}$$

Пусть $u \to 0$. Тогда из (23) получим уравнение

$$\ln Y = \frac{\ln X}{e^{\alpha \ln^{\beta} X}} \qquad (X, Y \ge 1), \tag{24}$$

которое может быть приведено к прямой

$$\ln \ln \frac{\ln X}{\ln Y} = \ln \alpha + \beta \ln \ln X. \tag{25}$$

Параметры α, β связного текста находятся путем построения по опытным значениям X_i, Y_i графика зависимости (25).

Параметры выборки можно рассчитать по трем величинам: X, Y, $Y_{m=1}$

$$\beta = \frac{1}{\ln \frac{\ln X}{\ln Y}}\left(1 - \frac{Y_{m=1}}{Y}\frac{\ln X}{\ln Y}\right), \tag{26}$$

$$\alpha = \frac{1}{\ln^{\beta} X} \ln \frac{\ln X}{\ln Y}. \tag{27}$$

Пусть далее $u = -I$. Тогда из (23) получим

$$\ln Y = \frac{\ln X}{\left(1 + \alpha \ln^{\beta} X\right)^2}, \tag{28}$$

откуда

$$\ln\left(\sqrt{\frac{\ln X}{\ln Y}} - 1\right) = \ln \alpha + \beta \ln \ln X. \tag{29}$$

Оценки параметров выборки равны

$$\beta = \frac{1 - \frac{Y_{m=1}}{Y}\frac{\ln X}{\ln Y}}{2\left(1 - \sqrt{\frac{\ln Y}{\ln X}}\right)}, \qquad (30)$$

$$\alpha = \frac{1}{\ln^{\beta} X}\left(\sqrt{\frac{\ln X}{\ln Y}} - 1\right). \qquad (31)$$

Достоинством формул (24), (28) является то, что они хорошо работают практически от начала координат до весьма больших значений X ($X = 10^7 \div 10^8$ словоупотреблений). Однако при дальнейшем увеличении X поведение этих кривых не соответствует поведению кривых роста новых событий.

Опыт показывает, что параметры выборки в приведенных выше формулах зависят как от типа текстов, на основе которых построена данная выборка, так и от ее объема. В связи с этим полученные формулы остаются справедливыми только в пределах заданной выборки, т.е. их нельзя использовать для экстраполяции кривой роста новых слов на выборку большего объема.

В то же время параметры текста не зависят от объема текста (при условии его лексической однородности), что позволяет прогнозировать объем словаря с ростом объема текста.

Проверим работу формул IIб и IУ (а, в) кривых роста новых слов в выборке. Восстановим кривую $y = f(x)$ на основе опытных данных по точным формулам, а также рассчитаем ее по приближенным формулам, при этом параметры выборки α_0, K , а также α, β определим по трем величинам: X, Y, $Y_{m=1}$.

Для восстановления кривой роста новых слов в выборке воспользуемся формулой В.М.Калинина (1964, с.247)

$$y = Y - \sum_{m \geq 1}\left(1 - \frac{x}{X}\right)^{m} Y_m , \qquad (32)$$

где Y - число разных слов в выборке объемом X словоупотреблений; Y_m - число слов с частотой m в выборке X ; y - ожидаемое среднее число разных слов в подвыборке произвольного объема x $(x < X)$.

Формулой (32) удобно пользоваться при $x/X \geq 0{,}1$.

При $x/X < 0{,}1$ целесообразно воспользоваться формулой (1), в которую вместо плотности $p(t)$ следует подставить относительные частоты слов $p_z^* = m_z^*/X$, где z - ранг слова в

частотном словаре. Тогда формула (I) примет вид

$$y = \int_0^n \left(1 - e^{-xp_z^*}\right) dz . \qquad (I')$$

Интегрирование осуществляется численным методом по формуле прямоугольников.

Таблица 2

Параметры выборок, рассчитанные по данным двух частотных словарей (ЧС)

Параметры	Пб (табл. I			IУа		IУв
	$u \approx I,25$	$u = 0,5$	$u \approx -0,25$	$u = -I$	$u \to 0$	$u = -I$
	ЧС немецкого языка					
α_0	0,02029	0,0I622	0,00996			
κ	0,0007485	0,00I287	0,002OI7			
α				0,04223	0,0057I9	0,002363
β				0,3005	I,374I	I,46635
	ЧС русского языка					
α_0	-0,00I633	-0,0I485	-0,03344			
κ	0,002935	0,004306	0,006I5			
α				0,0I45I	0,00I262	0,000466
β				0,4084	2,04I8	2,I8324

В таблице 3 (столбец 2) приведены результаты расчетов по точным формулам (I') и (32) на основе опытных данных частотного словаря немецкого языка и по приближенным формулам систем кривых роста Пб, IУа, IУв (столбцы 3-8). Объем выборки здесь равен X =I09I0777, объем словаря Y =258I73 и количество одноразовых слов $Y_{m=1}$ =I26862 (*Meiez* H., I964). Параметры выборки для каждой аппроксимирующей кривой приведены в табл. 2.

Аналогичные расчеты выполнены по данным "Частотного словаря русского языка" (под ред.Л.Н.Засориной, I977). Здесь X =I056382, Y =39268, $Y_{m=1}$ =I3379.

Из табл. 3 видно, что в первом случае все формулы достаточно точно описывают кривую роста новых слов в выборке, при этом наименьшую точность показала формула при $u \approx I,25$ (система Пб). Во втором случае менее точными оказались две формулы из той же системы кривых роста: при $u = 0,5$, $u \approx -0,25$.

Таблица 3

Кривые роста новых слов, восстановленные по двум ЧС и рассчитанные по формулам* систем кривых роста IIб, IVа, IVв

Объем выборки X	Объем словаря по ЧС Y	IIб			IVа	IVв	
		$u \approx 1{,}25$	$u = 0{,}5$	$u = -0{,}25$	$u = -1$	$u \to 0$	$u = -1$
ЧС немецкого языка							
I09I	625	600	6I9	646	603	626	636
2000	I0I9	979	I0I0	I054	I000	I0I9	I037
5000	2II8	2004	2066	2I56	2092	2082	2II6
I09II	3878	3609	37I6	3866	38I9	3737	3794
30000	-	7503	7699	7965	8009	7725	7828
I00000	I7800	I7II8	I7462	I79I2	I8207	I7485	I7663
300000	-	34735	352I5	358I5	36459	35220	35468
I09I078	77870	75440	75952	76545	77420	759I3	76I78
3273233	I408I0	I39063	I39377	I39670	I40230	I39293	I39474
I09I0777	258I73	258I67	258232	258I73	258I73	258083	258I77
ЧС русского языка							
I000	625	650	7I3	-	645	646	659
3000	I5I0	I534	I674	-	I57I	I526	I556
I0000	3637	36I8	3889	4274	3790	3604	3667
30000	7384	7296	7682	8I93	7670	7275	7374
I05638	I48I5	I4703	I5095	I5569	I52I5	I4683	I479I
2II276	20700	20597	20890	2I227	2I028	20582	20666
528I9I	30495	30455	30544	30645	30600	30450	30477
I056382	39268	39268	39269	39268	39268	39269	39269

* IIб

$u \approx 1{,}25$: $Y = X^{1/X^{\alpha/2}}$

$u = 0{,}5$: $Y = X^{1/(1+\frac{\alpha}{2}\ln X)}$

$u \approx -0{,}25$: $Y = X^{1/\sqrt{1+\alpha \ln X}}$

$(\alpha = \alpha_0 + \kappa \ln X)$

IV

а) $u = -1$: $y = \dfrac{x}{(1+\alpha x^{\beta})^2}$

в) $\begin{cases} u \to 0: \ln Y = \dfrac{\ln X}{e^{\alpha \ln^{\beta} X}} \\ u = -1: \ln Y = \dfrac{\ln X}{(1+\alpha \ln^{\beta} X)^2} \end{cases}$

Полученные результаты свидетельствуют о том, что параметр α изменяется от выборки к выборке. А это значит, что для более точного описания кривой роста новых слов в выборке необходимо использовать общие формулы (9), (I8), (23), каждая из которых (с учетом равенства (II) в случае формулы (9)) содержит три параметра. Для приближенных же расчетов удовлетворительные результаты дают рассмотренные выше формулы, имеющие всего по два параметра.

Возможность приведения этих формул к прямой оказалась весьма полезной для решения различных задач. Рассмотрим некоторые из них.

4.3. Оценка степени аналитичности языка

Показателем степени аналитичности языка принято считать коэффициент, равный отношению количества лексем к количеству словоформ частотного списка (Пиотровский Р.Г. и др., I962). Степень аналитичности языка тем выше, чем ближе этот показатель к единице.

Существенным недостатком этого показателя является его зависимость от объема текста. Знание аналитической зависимости между объемами словаря и текста позволяет по-другому измерять степень аналитичности языка.

Пусть кривая роста новых слов в тексте описывается формулой (I9). Величина параметра β зависит от выбора единицы подсчета количества разных слов, в качестве которой может быть принята словоформа (СЛ) или лексема (Л). При этом для одного и того же текста $\beta_л > \beta_{сл}$, в то время как $\alpha_л \approx \alpha_{сл}$. Последнее равенство позволяет ввести показатель степени аналитичности языка, который не зависит от объема текста:

$$\Delta\beta_a = \frac{v_л - v_{сл}}{\ln x}, \qquad (33)$$

где $v_л, v_{сл}$ рассчитываются по формуле

$$v = \ln\left(\sqrt{\frac{x}{y}} - 1\right). \qquad (34)$$

Чем ближе $\Delta\beta_a$ к нулю, тем выше степень аналитичности языка.

В табл.4 приведены значения $\Delta\beta_a$ для четырех языков, рассчитанные по формулам (33), (34) для текстов по электронике.

Полученные результаты позволяют осуществить переход от кривой роста новых словоформ в тексте к кривой роста новых лексем. Для этого достаточно воспользоваться формулами

$$\alpha_л = \alpha_{сл}\ ;\ \beta_л = \beta_{сл} + \Delta\beta_a\ .$$

Таблица 4

Степень аналитичности языков (тексты по электронике)

Язык, источник	Объем текста x	Объем словаря $y_{cл}$	Объем словаря $y_{л}$	$\Delta\beta_a$
Русский (Калинина,Е.А.,1968)	200894	21468	6826	0,0627
Румынский (Еван Л.И.,1966)	200000	14292	5708	0,0479
Французский (Кочеткова В.К., Скрелина Л.М., 1968)	100000	8108	4527	0,0336
Английский (Алексеев П.М., 1968)	200000	10582	7160	0,0202

4.4. Показатель степени связности слов в лексически однородном тексте

Так как в связном тексте лексико-грамматические связи накладывают определенные ограничения на сочетаемость слов, то естественно предположить (и это подтверждается опытными данными), что число разных слов в отрезке сплошного текста в среднем будет меньше, чем в случайно составленной выборке равного объема, взятой из достаточно большой совокупности лексически однородных текстов.

Эту разницу в объемах словарей можно использовать для оценки степени связности слов в тексте, причем, она оказывается независимой от объема текста.

Пусть для некоторого целого произведения из опыта известны объемы - всего текста X, словаря Y, одноразовых слов $Y_{m=1}$, а также несколько промежуточных точек $(x_i ; y_i)$ на кривой роста новых слов. По этим данным найдем параметры текста α_T, β_T путем построения графика зависимости (20). Для этого же произведения можно рассчитать параметры выборки α_B, β_B по формулам (21), (22). Эти параметры будут относиться к такой кривой роста, которая получилась бы при случайном отборе словоупотреблений из данного произведения и подсчете количества разных слов.

Многочисленные расчеты показали, что параметры α_T и α_B находятся в отношении

$$\frac{\alpha_T}{\alpha_B} \approx 2, \qquad (35)$$

которое может быть использовано в качестве показателя степени связности слов в тексте.

Параметры β_g и β_T связаны соотношением

$$\beta_T = \beta_g - \frac{\ell g 2}{\ell g X} \cdot \qquad (36)$$

формулы (35), (36) позволяют по параметрам выборки (которые легко вычислить по трем величинам X, Y, $Y_{m=1}$) найти параметры текста.

4.5. Оценка лексической близости двух связных текстов. Автоматическая классификация текстов

Пусть кривая роста новых слов в связном тексте описывается формулой (I9). Объединим два равных по объему текста с одинаковыми параметрами α, β и исследуем поведение обоих параметров этого объединенного текста. Рассмотрим два крайних случая.

Случай I. Параметры α, β объединенного текста равны соответствующим параметрам объединяемых текстов, а его словарь содержит такое же количество разных слов, сколько их было бы в каждом из объединяемых текстов при удвоенной его длине. В этом случае степень лексической близости двух текстов (будем измерять ее некоторым показателем ς) равна единице, а точка с координатами $(\ell n\, x_{12};\ v_{12\varsigma=1})$ лежит на прямой I (см.рис.I), поскольку

$$v_{12\varsigma=1} = \ell n\left(\sqrt{\frac{2x_1}{y_{12\varsigma=1}}} - 1\right) = \ell n\alpha + \beta \ell n 2x_1 = v_1 + \beta \ell n 2 .$$

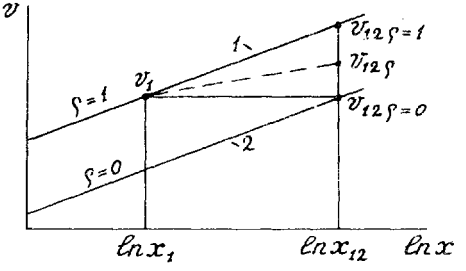

Рис. I. Графическое определение показателя ς

Случай 2. Словари обоих текстов не содержат общих слов. Это значит, что степень лексической близости таких текстов равна нулю ($\varsigma = 0$). В этом случае объем словаря объединенного текста равен удвоенному объему словаря одного из объединяемых текстов, а величина

$$v_{12\varsigma=0} = \ln\left|\sqrt{\frac{2x_i}{2y_i}} - 1\right| = v_i = \ln\alpha + \beta \ln x_i ,$$

т.е. прямая 2 параллельна прямой I и расположена ниже ее на расстоянии $\beta \ln 2$.

Из рис. I видно, что величина $v_{12\varsigma}$, характеризующая объем словаря реального объединенного текста, ограничена $v_{12\varsigma=0} < v_{12\varsigma} < v_{12\varsigma=1}$. Следовательно, показатель ς можно измерять отношением

$$\varsigma = \frac{v_{12\varsigma} - v_{12\varsigma=0}}{v_{12\varsigma=1} - v_{12\varsigma=0}} = \frac{v_{12\varsigma} - v_{12\varsigma=0}}{\beta \ln 2} . \qquad (37)$$

Из условия параллельности прямых I и 2 следует, что ς не зависит от объема текстов x_i, но оба они должны быть равными между собой.

Отметим, что для двух сравниваемых текстов отношение $y_{12\varsigma=0} / y_{12\varsigma=1}$ не может превосходить определенного значения, которое при $0 < x_1 < \infty$ заключено на интервале $1 < y_{12\varsigma=0} / y_{12\varsigma=1} < 4^{\beta}$. Например, для "Частотного словаря русского языка" при β = 0,4084 (табл.2, столбец 5) это отношение не должно превышать величины 4^{β} = I,76I5. Действительно, при x_1 = 528I9I по данным табл. 3 (столбец 2) имеем: 30495 · 2/39268 = I,5532.

Параметр β, входящий в формулу (37), может быть надежно определен лишь по тексту достаточно большого объема. В случае коротких текстов показатель ς можно оценить проще, если объем объединенного текста принять равным $x_{12} = x_1/2 + x_2/2$. Тогда для вычисления показателя ς достаточно будет найти значения следующих величин:

$$y_{12\varsigma=0} = y_1(x_1/2) + y_2(x_2/2); \; y_{12\varsigma} = y(x_{12});$$

$$y_{12\varsigma=1} = \frac{1}{2}\left[y_1(x_1) + y_2(x_2)\right],$$

где $y_1(x_1/2)$ - количество разных слов среди половины словоупотреблений первого текста (например, стоящих только на четных местах); $y_2(x_2/2)$ - то же для второго текста; $y(x_{12})$ - количество разных слов среди половины словоупотреблений первого и второго текстов; $y_1(x_1)$, $y_2(x_2)$ - количество разных слов соответственно в первом и втором текстах.

При небольших размерах текстов ($x < 40000$) кривая роста новых слов достаточно точно описывается формулой

$$y = \frac{x}{1 + \alpha x^{\beta}} ,$$

которая следует из (I8) при $u \to -\infty, \alpha u > 0$. Ее можно привести к виду

$$v = \ln\left(\frac{x}{y} - 1\right) = \ln\alpha + \beta \ln x . \qquad (38)$$

Для определения показателя $\int$ (см. формулу (37)) необходимо по формуле (38) вычислить значения $v_{12\int}$ при заданных x_{12} и $y_{12\int}$. При этом получим оценку $\int$, не зависящую от объема сравниваемых текстов (но оба текста должны быть равными между собой).

Все операции по вычислению $\int$ легко автоматизировать. Задавая пороговое значение этого показателя, можно классифицировать тексты (документы) по тематическим группам, включая в одну группу близкие по содержанию тексты.

4.6. Система IУа кривых роста информационных потоков

Одной из таких кривых, принадлежащих данной системе, является кривая роста (во времени x) числа названий книг и брошюр. Ее свойства отличаются от свойств кривых роста новых слов. При этом, как показывает анализ опытных данных, рост числа названий книг и брошюр, заявок на изобретения и т.д. может быть либо линейным, либо близким к экспоненте. Следовательно, общее уравнение для описания этих кривых роста может быть записано в виде (при $\gamma = I$, $\alpha < 0$, $0 < u < I$)

$$y = y_0 \left(1 - \alpha u x^{\beta}\right)^{\frac{1}{u} - 1}, \qquad (39)$$

где y_0 - начальное значение y при $x = 0$.

Для нахождения параметров α, β, u преобразуем (39) к виду

$$\ln\left[\left(\frac{y}{y_0}\right)^{\frac{u}{1-u}} - 1\right] = \ln\alpha u + \beta \ln x . \qquad (40)$$

Значение параметра u должно быть подобрано таким, чтобы график зависимости (40), построенный по опытным значениям y, x, представлял собой прямую. Оценки параметров α, β могут быть определены по графику полученной прямой.

В случае, если ни при каких значениях параметра u $(0 < u < I)$ прямая не получается, то это свидетельствует о том, что формула (39) в данном случае не работает и следует рассмотреть кривую другого типа, например,

$$y = y_0 e^{\alpha x^{\beta}} . \qquad (4I)$$

Оценки параметров α, β находятся на основании уравнения прямой

$$\ln\ln\frac{y}{y_0} = \ln\alpha + \beta\ln x \qquad (42)$$

при $\alpha > 0$, либо прямой

$$\ln\ln\frac{y_0}{y} = \ln\alpha + \beta\ln x \qquad (43)$$

при $\alpha < 0$. Исследования показывают, что кривая роста (41) при $\beta < 0$ имеет горизонтальную асимптоту $y = y_0$, которую также следует оценить по опытным данным. Для этого представим формулу (41) в виде

$$y_{nx} = a y_x^b , \qquad (44)$$

где y_{nx} может обозначать число наступивших разных событий в выборке объемом nx (за время nx). Параметры a, b находятся из уравнения прямой

$$\ln y_{nx} = \ln a + b\ln y_x . \qquad (45)$$

Тогда оценки параметров в формуле (41) будут равны

$$y_0 = a^{1/(1-b)} \quad (b \neq 1) , \qquad (46)$$

$$\beta = \frac{\ln b}{\ln n} , \qquad (47)$$

$$\alpha = \frac{1}{s}\sum_{i=1}^{s}\frac{1}{x_i^\beta}\ln\frac{y_i}{y_0} . \qquad (48)$$

Если $b = 1$, то кривая роста задается формулой $y = \alpha x^\beta$.

Формулы (41)-(45) хорошо описывают кривые роста числа названий книг и брошюр, кривые роста производительности труда и произведенного национального дохода (Нешитой В.В., 1984) и некоторые другие кривые. Важной характеристикой этих кривых является темп роста

$$q_i = \frac{y_i}{y_{i-1}} = e^{\alpha(x_i^\beta - x_{i-1}^\beta)} . \qquad (49)$$

Из (49) следует, что при $\beta = 1$ $q_i = e^\alpha = const$, т.е. темп роста не зависит от времени x_i и равен темпу роста показательной функции $y = e^{\alpha x}$. При $\beta > 1$ величина q_i с ростом x_i растет, а при $\beta < 1$ – уменьшается, т.е. параметр β является показателем ускорения или замедления темпа роста кривой (41). При $x_1 = 1$,

$x_0 = 0$ из (49) имеем $q_1 = y_1/y_0 = e^{\alpha}$, откуда $\alpha = \ln(y_1/y_0) = \ln q_1$, т.е. величина α представляет собой натуральный логарифм темпа роста кривой на начальном отрезке времени, равном единице. Поскольку $q_1 = e^{\alpha} = 1 + \frac{\alpha}{1!} + \frac{\alpha^2}{2!} + \ldots$, то при малых α ($\alpha < 0{,}1$) имеем $q_1 \approx 1 + \alpha$, откуда $\alpha \approx q_1 - 1$.

Проверим работу формул (41), (44) на практическом примере. Воспользовавшись опытными данными (Михайлов А.И. и др., 1976, с.212, 213), приведенными в табл.5, столбцы 1,3, найдем параметры выравнивающей кривой роста числа названий книг и брошюр, изданных в 16 странах за 1958-1972 г.г. Пусть эмпирическая кривая роста описывается формулой (41), которую представим в виде

$$\frac{y}{y_0} = e^{\alpha x^{\beta}}, \qquad (50)$$

где y_0 = 216696 (количество книг и брошюр на конец 1958 г.). Приведем ее к уравнению прямой

$$\left(\ln \frac{y}{y_0}\right)_{nx} = b\left(\ln \frac{y}{y_0}\right)_x. \qquad (51)$$

Таблица 5

Число названий книг и брошюр (y), изданных в 1958-1972 г.г. в 16 странах (Михайлов А.И. и др.,1976)

Год	x	y	$\frac{y}{y_0}$	$\ln \frac{y}{y_0}$	$\left(\frac{y}{y_0}\right)^{расч.}$
1958	0	216696	1	0	1
1959	1	224201	1,0346	0,0341	1,0378
1960	2	237837	1,0976	0,0931	1,0824
1961	3	245382	1,1324	0,1243	1,1314
1962	4	257619	1,1889	0,1730	1,1843
1963	5	269942	1,2457	0,2197	1,2410
1964	6	282892	1,3055	0,2666	1,3016
1965	7	288422	1,3310	0,2859	1,3662
1966	8	323388	1,4924	0,4004	1,4350
1967	9	333848	1,5406	0,4322	1,5081
1968	10	336975	1,5551	0,4415	1,5858
1969	11	347699	1,6045	0,4728	1,6682
1970	12	385570	1,7793	0,5762	1,7557
1971	13	390247	1,8009	0,5883	1,8486
1972	14	396661	1,8305	0,6046	1,9471

Если построить по эмпирическим значениям величин $\ln(y/y_0)$ график зависимости (5I) при $n = 2$, т.е при всех возможных парах значений $(x, 2x)$: I,2; 2,4; 3,6;...; 7,I4, - то убедимся, что точки действительно рассеиваются вдоль прямой (5I), проходящей через начало координат. Ее угловой коэффициент $b = 2,I3537$. Следовательно, согласно (47), параметр $\beta = I,0945$. Параметр α, вычисленный по формуле (43), равен $\alpha = 0,03709$.

Таким образом, выравнивающая кривая роста может быть записана в виде

$$y = 216696\, e^{0,03709(x-1958)^{1,0945}}.$$

Поскольку параметр $\beta > I$, то темп роста числа названий книг и брошюр со временем увеличивается.

Л И Т Е Р А Т У Р А

Алексеев П.М. Частотный словарь английского подъязыка электроники // Статистика речи. - Л., I968. - с. I5I-I6I.

Ешан Л.И. Опыт статистического описания научно-технического стиля румынского языка (на материале текстов по радиоэлектронике): Автореф.дис... канд.филол.наук. - Л., I986. - I6 с.

Калинин В.М. Некоторые статистические законы математической лингвистики // Проблемы кибернетики. Вып. II. - М., I964 - с. 246-255.

Калинина Е.А. Изучение лексико-статистических закономерностей на основе вероятностной модели // Статистика речи. - Л.:Наука, I968. - с. 64-I07.

Кочеткова В.К., Скрелина Л.М. Частотный словарь французского подъязыка электроники// Статистика речи. - Л., I968. - с. I62-I70.

Михайлов А.И., Чёрный А.И., Гиляревский Р.С. Научные коммуникации и информатика. - М.: Наука, I976. - 436 с.

Нешитой В.В. Система непрерывных распределений как экономико-математическая модель// Проблемы макроэконометрического моделирования и прогнозирования: Тезисы докладов респ.конф./ АН Латв.ССР; ин-т эк-ки. - Рига: Зинатне,I984. - с. 237-239.

Нешитой В.В. Исследование ранговых распределений // НТИ. Сер.2. - I985. - № 2. - с. I6-20.

Нешитой В.В. Ранжирование слов по степени семантической нагрузки// НТИ. Сер.2. - I986. - № 4. - с. 20-25.

Пиотровский Р.Г., Алексеев П.М., Чернядьева Е.А. Статистика речи и закономерности языка// Тез.докл.межвуз.конф. на тему "Язык и речь" (27 ноября - 7 декабря)/ ЛГПИИ. - М.,I962. - с. 57-59.

Частотный словарь русского языка. / Под ред.Л.Н.Засориной. - М.: Русский язык, I977. - 935 с.

Meier H. Deutsche Sprachstatistik. - Hildesheim, 1964.

MATHEMATICAL MODELS OF THE DICTIONARY
EXPANSION AND DATA FLOWS

V.V. Neshitoi

S u m m a r y

Four systems of the mathematical models for description of different growth curves are proposed, methods for parameter evaluation are given.

Indices of the analytical character of the language, word cohesion in a lexically uniform text and lexical similarity of two connected texts are introduced, these indices being independent of the text size.

МНОГОМЕРНАЯ ВЕРБАЛЬНАЯ СТРУКТУРА СТЕРЕОТИПОВ В РАЗНЫХ СТАДИЯХ ИБС

В.А. Отлыгин

Широта подходов и точек зрения, обнаруживаемых в квантитативной лингвистике, определяется не только дифференциацией наук. В большей степени это связано с интерпретацией получаемых данных человеческим сознанием. Последнее определение весьма нечетко и только в определенном смысле может быть использовано в "нечеткой математике". Видимо, существует определенное соответствие между степенью неопределяемости нашего сознания и количеством предлагаемых теорий для его анализа, особенно в отношении вербализации его и порождаемого в этом случае текста.

В квантитативной лингвистике, при всем многообразии методов исследования, преобладает статистический подход, точнее, статистические информационные принципы исследования (Гачечиладзе Т.Г., Манджапарашвили Т.В., 1987; Тулдава Ю.А., 1987; Черии К., 1972). Этот подход распространяется и на психолингвистические разработки. В этом случае также рассматривается применимость (расширение) закона Ципфа на реакцию человека в ответ на вербальный раздражитель (Тулдава Ю.А., 1987; Черии К., 1972; Борода М.Г., Поликарпов А. А., 1987).

В случае большого числа наблюдений такой подход правомерен, электроэнцефалографические данные могут интерпретироваться таким образом (Тулдава Ю.А., 1987), на небольших выборках эти закономерности не всегда прослеживаются (Шубина Т.И., Грачева Н.И. и др., 1987), хотя различные параметры вербальной реакции и реакции на вербальный раздражитель, как правило (без соблюдения условий применения корреляционного анализа) коррелируют. Особенно это заметно при проведении психоэмоциональных проб (Шубина Т.И., Грачева Н. И. и др., 1987).

Заслуживает внимания другое обстоятельство. Психолингвистический подход в квантитативной лингвистике в конечном счете сводится к нахождению соответствия с типологией личности. Последние нормативы являются конвенциальными, что в итоге образует своего рода порочный круг. Выбраться из него также затруднительно и специалистам психиатрического профиля. Для психиатра использование количественных методов ана-

лиза часто является некорректным, хотя корреляционный анализ также, как и в других медицинских исследованиях, используется довольно широко (Вальдман А.В., Александровский Ю.А., 1987).

Альтернативой такому положению может служить гипотеза о своеобразной системной конкуренции анализаторных систем (Отлыгин В.А., Еремин С.В., 1988). Суть ее можно свести к тому положению, что восприятие и воспроизведение речи на уровне т.н. "внутреннего говорения" организовано на принципах работы анализаторов, например, слухового и зрительного (до 80 % всего объема информации) (Черии К., 1972). Помимо разноуровневого анализа слов, для первого характерны те способы анализа, где используются представления о потенциале, временные параметры; для второго - двумерность восприятия (Рамачандран В.С., 1988), образование по различным принципам третьей координаты (например, по известному принципу Анжера), различие на границах изображения, образов. В этом случае использование электроэнцефало- и кардиографических и других приемов в поиске квантов текста не является универсальным; возможность анализа ограничивается мощностью слухового анализатора и способностью взаимодействия с результатами деятельности других анализаторных систем. Заметим, что, видимо, функциональная ассиметрия головного мозга (Брагина Н.Н., Доброхотова Т.А., 1988; Восприятие речи, 1988) определяется не только анатомическими особенностями - более вероятно, что они вторичны, и реализуются довольно сложными взаимодействиями (Судаков К.В. (ред.), 1987).

Если обсуждаемая гипотеза правомерна, то на определенном этапе исследования представляется необходимым проанализировать систему, в которой представлены социально-биологические предпочтения (мотивы) личности и их реализация на вербальном уровне. Текст анализировался по параметрам: качество высказываний, "вес" объекта и действия, "вес" узкоденотативных знаков. Под качеством высказываний понимается отношение слов в ответе к количеству слов в полном предложении, которое могло быть в тексте. В каждом высказывании имелись слова (группа слов), относящиеся к объекту, либо к действию. Идентификация проводилась в соответствии с работой (Дудников А.В., 1983). Наконец, выбор такого параметра, как узкоденотативные знаки определялся не только тем, что они являются "метками" образа и определяются, например, при дефиците кислорода или состоянием, сходным с аффективным (Спивак

Д.Л., 1986), но и тем, что при нарушении взаимодействии систем анализаторов уже на уровне переработки, например, при интерференции носителей тех или иных черт анализаторов. Здесь предполагается, что качество высказываний и узкоденотативные знаки показывают слабую дифференциацию или отсутствие ее на уровне "внутренней речи". Иначе говоря, эти знаки проходят ее "насквозь", без анализа.

Так как идентификация состояния пациента и анализ текста им порождаемого проводился в отношении социально-биологически значимых для него проблем в условиях конфликтной ситуации, то выбирались пациенты сравнительно одной социальной ориентации из т.н. группы "риска" к ишемической болезни сердца; из группы, образованной пациентами после перенесенного инфаркта миокарда (на 3-ей неделе); и - из группы пациентов, начавших работу (т.н. рабочая неделя) после инфаркта миокарда под наблюдением врача (через 3 месяца). Возраст - 45 - 50 лет, в каждой группе - 10 человек.

Выбор патологии становится обоснованным, если учесть широкую ее распространенность среди населения, а так же то, что в прединфарктный период преобладает отрицание возможности заболеть (анозогнозия), после инфаркта - субдепрессивный симптомокомплекс.

Никаких предварительных инструкций перед исследованием, в отличии от принятого тестирования, по нашей методике не давалось. Результаты представлены в виде компонентов вектора (в %%), углов между ними. По дополнительной программе "Шаман" рассчитывались также углы между векторными слоями.

Здесь необходимы некоторые пояснения. Помимо указанных параметров исследований в таблицу внесены результаты расчета времени на вербальный раздражитель. При этом компоненты временного вектора и векторного слоя отражают не только время задержки в ответе, но соотносительное между проблемами время реагирования. Впрочем, все параметры даются как бикомпаративные, т.е. результаты получены путем попарных оценок (Отлыгин В.А., Клементьев А.А. и др., 1988). Целью такого анализа явилась, в частности, попытка (альтернативная) избежать ортогональности, как это имеет место практически во всех тестах, в представлении результатов исследования. Более того, нет никаких убедительных данных, что параметры изучения личности обязательно должны располагаться относительно друг друга под углом в 90 градусов, как это принято в тестировании. С учетом человеческого двумерного восприятия ортого-

нальность не более, чем попытка деформировать представления о личностных характеристиках в привычное пространство, например, з-х мерное.

Такие же исследования проводились с экспертами. Необходимо указать, что здесь касаемся довольно сложной проблемы эксперта. Дело в том, что сама по себе экспертная оценка не является чем-то новым, но в психологических и психиатрических исследованиях эксперт является своего рода верховным судьей, и его личностные характеристики, а также его состояние не учитываются. Это традиционный, но не правомерный подход, скорее вынужденный, т.к. мы не имеем некого существа, способного бесстрастно и несообразно своего состояния оценивать ситуацию. Именно поэтому в методику исследования привносятся параметр-угол между мнением и другими оценками пациента и эксперта. Эта величина характеризует, помимо всего прочего, степень вживаемости эксперта в конфликтную ситуацию, степень его "понимания" состояния пациента.

Обращает на себя внимание, что оценка своего здоровья (I-я проблема) варьирует в разных группах, и в I-ой - значимость ее минимальна в оценке пациента. Значимость работы (2-ая проблема) резко возрастает перед выходом на работу; в этой группе отношения с женой (3-я проблема) менее значима. Оценка своего "Я" (6-я проблема) минимальна в последней группе, но время, затрачиваемое при ее обсуждении минимально в I-ой группе; в других группах - время сопоставимо по величине не только друг с другом, но и при обсуждении других проблем. Качество высказывания снижается только во второй группе; в этой и 3-ей - описания действия более точно, чем объекта: выбор слов для программы действия таков, что всегда присутствуют элементы долженствования, императивности.

Компоненты вектора УI показывают, что наибольшее количество узкоденотативных знаков обнаруживается при обсуждении проблемы "работы" (I-я группа); по другим проблемам - доля их примерно одинакова. В других группах дифференциация такова, что при обсуждении финансовых проблем и значимости "Я" (5,6 проблемы) они малоупотребимы. Иначе говоря, эти проблемы в вербальном отношении определены. Другие проблемы во 2-ой группе имеют соотносительно столько узкоденотативных знаков (в высказываниях), что проявляется зависимость от значимости соответствующей проблемы. В определенном смысле для пациента эти стереотипы столь жестки, что для их идентификации ему достаточно только обозначить их - "метка" работает

Таблица I

Компоненты вектора: мнений (I), соотносительного времени (П), качества высказываний (Ш), объекта (IУ), действия (У), "веса" узкоденотативных знаков (УI), (в %%), испытуемых (*)

№ проблемы	Группы: 1. Риск к ИБС — слои						2. после инфаркта миокарда — слои						3. перед выходом на работу — слои					
	I	П	Ш	IУ	У	УI	I	П	Ш	IУ	У	УI	I	П	Ш	IУ	У	УI
1.	$\underline{5}_{34}$	$\underline{22}_{30}$	$\underline{20}_{16}$	$\underline{29}_{34}$	$\underline{32}_{33}$	$\underline{10}_{48}$	$\underline{30}_{46}$	$\underline{15}_{25}$	$\underline{31}_{26}$	$\underline{39}_{34}$	$\underline{42}_{35}$	$\underline{36}_{51}$	$\underline{22}_{46}$	$\underline{15}_{29}$	$\underline{9}_{18}$	$\underline{34}_{34}$	$\underline{36}_{34}$	$\underline{23}_{49}$
2.	$\underline{25}_{28}$	$\underline{16}_{22}$	$\underline{19}_{21}$	$\underline{20}_{25}$	$\underline{22}_{27}$	$\underline{34}_{29}$	$\underline{23}_{27}$	$\underline{21}_{25}$	$\underline{25}_{24}$	$\underline{21}_{25}$	$\underline{23}_{28}$	$\underline{17}_{27}$	$\underline{51}_{26}$	$\underline{12}_{26}$	$\underline{19}_{17}$	$\underline{28}_{27}$	$\underline{25}_{29}$	$\underline{19}_{28}$
3.	$\underline{39}_{20}$	$\underline{22}_{20}$	$\underline{12}_{15}$	$\underline{16}_{15}$	$\underline{15}_{19}$	$\underline{14}_{10}$	$\underline{15}_{13}$	$\underline{20}_{20}$	$\underline{22}_{21}$	$\underline{19}_{17}$	$\underline{18}_{19}$	$\underline{24}$ 9	$\underline{13}_{11}$	$\underline{22}_{21}$	$\underline{13}_{15}$	$\underline{14}_{15}$	$\underline{17}_{19}$	$\underline{26}$ 9
4.	$\underline{9}$ 8	$\underline{11}_{13}$	$\underline{17}_{14}$	$\underline{5}_{11}$	$\underline{4}_{10}$	$\underline{11}$ 7	$\underline{5}$ 5	$\underline{14}_{15}$	$\underline{6}$ 8	$\underline{11}_{11}$	$\underline{8}$ 8	$\underline{12}$ 7	$\underline{2}$ 5	$\underline{12}_{13}$	$\underline{20}_{14}$	$\underline{14}_{12}$	$\underline{12}$ 8	$\underline{19}$ 7
5.	$\underline{3}$ 3	$\underline{20}_{10}$	$\underline{10}_{11}$	$\underline{15}$ 7	$\underline{13}$ 7	$\underline{17}$ 4	$\underline{5}$ 5	$\underline{13}$ 9	$\underline{7}_{10}$	$\underline{5}$ 7	$\underline{5}$ 6	$\underline{7}$ 4	$\underline{3}$ 3	$\underline{22}$ 7	$\underline{28}_{20}$	$\underline{6}$ 3	$\underline{7}$ 6	$\underline{8}$ 7
6.	$\underline{16}$ 8	$\underline{8}$ 4	$\underline{20}_{18}$	$\underline{15}$ 5	$\underline{10}$ 4	$\underline{13}$ 2	$\underline{20}$ 4	$\underline{17}$ 6	$\underline{8}_{11}$	$\underline{5}$ 6	$\underline{4}$ 4	$\underline{4}$ 2	$\underline{9}$ 9	$\underline{17}$ 4	$\underline{11}_{16}$	$\underline{4}$ 9	$\underline{3}$ 4	$\underline{5}$ 2

* - на пол-интервала ниже приводятся величины компонентов вектора эксперта (расчетный вектор); величины компонентов "прямого" вектора для исследуемого и эксперта - не приводятся: сопоставление величин расчетного и прямого векторов обычно приводятся для определения инстинных (или неосознаваемых) мотивов (Отлыгин В.А. и др., 1988).

Таблица 2

Величины углов между векторами и слоями векторов

	Группы: 1. Риск к ИБС — слои						2. после инфаркта миокарда — слои						3. перед выходом на работу — слои					
	I	П	Ш	IУ	У	УI	I	П	Ш	IУ	У	IУ	I	П	Ш	IУ	У	УI
пац./экс.	41	19	-	18	15	49	20	16	-	-	-	23	27	26	-	18	15	49
слои век.	52	23	22	52	52		33	38	32	36	36		38	17	49	13	-	

** - "~" величина угла не превышает 10 градусов,
величины углов между другими слоями не приводятся, но обсуждаются в тексте.

автоматически. Одновременно из этого следует, что какие-либо ситуации, касающиеся этих стереотипов и требующих обсуждения должны вызывать определенный дискомфорт. Иначе говоря, включение каких-либо систем для вербализации - во "внутренней речи", сообразие с признаками работы анализирующих систем не требуется для такого пациента (понятно, и личности).

Следует отметить, что Ш-й и УI-ой слои в определенной мере связаны; предполагаем, что это разные этапы, на которых замедлилось формирование стереотипов. Правда, для Ш-го слоя, видимо, характерна анафорно-катафорная связь (Падучева Е.В., 1988) и "тема-рема"-тические отношения, но формализация их еще разрабатывается. Можно продолжить этот анализ с психиатрической точки зрения: ухудшение качества высказываний и, далее, преобладание узкоденотативных знаков - своего рода этапы сужения восприятия (ограничения сознания на психотравмирующей ситуации и депримированность) и мышления. Возможно, ограниченно прав автор (Выготский Л.С.,1934), говоря, что "значение слова ... есть единство слова и мысли".

Наибольшее несоответствие, т.е. величина угла между слоями, отмечается между компонентами вектора мнения и временем реагирования (I гр.) и между слоями Ш и IУ в I и 3 группах. Иначе говоря, соответствие отмечается только по малому перечню проблем. Анализ этого явления выходит за рамки данного сообщения, но, очевидно, нельзя исключить влияния других проблем и факторов, например, ролевой деформации в поведении и оценках состояния.

Ранее было показано (Отлыгин В.А., Клементьев А.А. и др., 1988), что при глубокой депрессии величины компонентов векторов практически одинаковы; по данным таблицы I можно считать, что у исследуемых во всех группах наблюдаются признаки субдепрессии.

Следует указать, что "понимание" проблем пациента экспертом достигается во 2-ой группе; в I-ой группе понимание и вживаемость в состояние пациента явно недостаточно. Полное непонимание - при расхождении векторов на 90 градусов. Выражение, что суждения о чем-либо перпендикулярно, в данном случае обретает не только образный смысл, а может объективизироваться. Во всех группах отмечается расхождение в оценках пациента и эксперта в структуре высказываний. Попытки избежать влияния на ответы и монологи пациента приве-

ли к значительному расхождению в величинах компонентов вектора в УI слое во всех группах в отношении I проблемы. Здесь можно говорить о ролевых, точнее - профессиональных деформациях. Если учесть, что рассчетный вектор характеризует внутренние (близкие к истинным) мотивы и стереотипы, то по данным этих таблиц правомерно считать, что эксперт является человеком своей - в данном случае врачебной - культуры.

Между всеми группами (и векторами) отмечается расхождение: I - 2 - 43 град., I-3 - 34 град., 2-3 - 30 град. В контексте сообщения такое различие вполне обосновано. Дело не только в статистически достоверном и фактическом различии, но и в том, что значимость стереотипов и их границы варьирует в разных группах. Компоненты векторов во П слое показывают, что переработка информации у пациентов значительно различается; можно принять, что, если в формировании и работе памяти реализуется голографический принцип (Судаков К.В. (ред.), I987), то его проявление в чистом виде по этим данным - не более, чем частный случай, в реальности дело сложнее. Мы наблюдали при логически точной оценке (нарушении принципа нетранзитивности, а он более реален, чем логика) матрицы оценок близки к вырождению, а критерии статистической оценки мало применимы. Видимо, слово несет в себе признаки функционирования всех анализаторных систем.

Л И Т Е Р А Т У Р А

Борода М.Г., Поликарпов А.А. Семиотические универсалии и их роль в организации и развитии музыкального и естественного языка// Квантитативные аспекты системной организации текста. - Тбилиси, I987.

Брагина Н.Н., Доброхотова Т.А. Функциональные ассиметрии человека. - М.: Медицина, I988.

Вальдман А.В., Александровский Ю.А. Психофармакотерапия невротических расстройств. - М.: Медицина, I987.

Восприятие речи. Вопросы функциональной ассиметрии мозга. - Л.: Наука, I988.

Выготский Л.С. Мышление и речь. - ОГИЗ-СОЦЭГИЗ, I934.

Гачечиладзе Т.Г., Манджапарашвили Т.В. Нечеткие множества и статистическая лингвистика // Квантитативные аспекты системной организации текста. - Тбилиси, I987.

Дудников А.В. Русский язык. - М.: Просвещение, I983.

Отлыгин В.А., Еремин С.В. Оценка эффективности психоэмоциональных проб для клинической практики // Актуальные вопросы реабилитации. - М.: Прейскурантиздат, 1988.

Отлыгин В.А., Клементьев А.А., Ползухина М.Э., Шифрин В.Л., Бычкова Т.Г. Некоторые количественные характеристики ментальной дезадаптаций. Деп. Д-15929-05.08.1988 ВНИИМИ МЗ СССР.

Падучева Е.В. Снова анафора и коферентность // Вопросы кибернетики. - М.: АН СССР, 1988.

Рамачандран В.С. Светотени и восприятие формы // В мире науки. № 10, 1988.

Спивак Д.Л. Лингвистика измененных состояний сознания. - Л.: Наука, 1986.

Судаков К.В. (ред.) Функциональные системы организма. - М.: Медицина, 1987.

Тулдава Ю.А. К вопросу о классификации и интерпретации лингвистических распределений // Ученые записки Тартуского университета, вып. 774. - Тарту, 1987.

Черии К. Человек и информация. - М.: Связь, 1972.

Шубина Т.И., Грачева Н.И., Храмелашвили В.В., Ревенко В.Н., Гросу А.В., Сидоренко Б.А. Чувствительность к психоэмоциональной пробе и особенности психологического статуса больных ишемической болезнью сердца // Кардиология. Т. XXУП, 6, 1987.

MULTIDIMENSIONAL VERBAL STRUCTURE OF STEREOTYPES AT DIFFERENT STAGES OF CHD

Vladimir A. Otlygin

S u m m a r y

Statements of patients with the most common pathology have been analyzed at different stages of its manifestation. The values of vector components were used to determine the sociobiological preferences and were correlated with the quality of statements, the time of response, the "weight" of the object, the "weight of the action" and narrow-denotative symbols in the statements. The generated text proved to be formed on the basis of stereotypes with different verbal structures; from the value of the angles between vectors and vector layers was determined the degree of the studied parameters divergence from the text. A number of hypotheses regarding the text formation and the principles of its organization are discussed. From the expert's estimate were determined his "understanding" of the patient's state, the role of the expert's professional deformation in estimating the patient's text.

О СИСТЕМНОМ СООТНОШЕНИИ КРАТКОГО И СРЕДНЕГО ТОЛКОВЫХ СЛОВАРЕЙ РУССКОГО ЯЗЫКА

А.А. Поликарпов, О.С. Фрюкова

I. Исследования по типологии толковых словарей, начатые С.И. Ожеговым (1974), по ряду причин не получили достаточного развития[+]. Одной из главных причин является то, что в современной лингвистике не осознан принципиальный факт соответствия разных типов словарей разным областям языкового сознания общества, не осознан факт специфического отображения в типологии словарей структуры объективного лексического пространства языка как коммуникативно-социального явления.

Как и всякое отображение, словарь несет в себе какие-то черты непоследовательностей или неточностей, связанных с тем, что реальность отображают не автоматы, а люди. Критика словарей, как правило, сосредотачивается на этих непоследовательностях и неточностях. Однако за "лексикографическими деревьями" надо видеть и "языковой лес", видеть структуру лексики, системную организацию определенных участков совокупного лексикона общества, отображаемых в толковых словарях.

В настоящей работе проводится попытка объективного анализа состава толковых словарей на основе сопоставления семантических и стилистических характеристик лексических единиц краткого словаря – "Словаря русского языка" С.И Ожегова (9-е изд., 1972 г., под ред. Н.Ю. Шведовой) и словаря среднего типа – "Словаря русского языка" в 4-х томах (2-е изд., 1981–1984 г.г., под ред. А.П. Евгеньевой), проводится попытка разделения систематических и случайных фактов, отражаемых в них. Эта статья является частью общей программы сопоставительного исследования словарей разного типа в одном языке и одного типа в разных языках.

2. Типология толковых словарей русского языка впервые в принципиальном виде представлена в работе С.И. Ожегова "О трех типах толковых словарей русского языка" (1952). С.И.

[+] Вопросы общей типологии словарей, в пределах которой толковые получают свое определение на фоне идеологических, исторических, энциклопедических и т.п. рассматриваются в работах (Щерба, 1958; **Malkiel**, 1959; **Sebeok**, 1962; Апажев, 1971; **Rey-Debove**, 1971; Новиков, 1973; Копецкий, 1972; Трофимкина, 1972; Денисов, 1974).

Ожегов писал: "Практически созданы и теоретически намечены три основных типа нормативных общих словарей русского языка: большой, представляющий современный литературный язык в широкой исторической перспективе; средний, с детальной разработкой исторически оправданного стилистического многообразия современного литературного языка, и, наконец, краткий, популярного типа, стремящийся к активной нормализации современной литературной речи" (Ожегов, 1974, с. 165-166).

Задачи формирования словника словарей разных типов, по С.И. Ожегову, сводятся к следующему: "Большой словарь включает в себя активный и пассивный запас лексики современного общенародного литературного языка и вышедшую из употребления, но характерную для развития словарного состава лексику. Средний словарь включает в себя активный и пассивный запас лексики современного языка. Краткий словарь обнимает активный запас современной лексики с привлечением той лексики пассивного запаса, которая необходима с той или иной точки зрения для характеристики современного языка" (Ожегов, 1974, с. 173).

Как обнаруживает специальный анализ, тождественные или близкие к этому членению типы толковых словарей вырабатываются в любой развитой лексикографической традиции. По всей видимости, это отражает реальное коммуникативно-социальное членение литературного лексикона данного народа, которое для удовлетворения социальной потребности в дифференцированном отражении каждого из пластов организации нормативного лексикона общества требует построения отдельных словарей.

Современная трактовка трех основных пластов литературного словаря, получаемая на основе специального словарного тестирования носителей языка (Поликарпова, Поликарпов, 1987), заключается в том, что "активный запас" современной лексики, отражаемой в кратком словаре, "стремящемся к активной нормализации", хорошо соответствует пересечению всех активных (используемых продуктивно) индивидуальных словарей. "Активный и пассивный запас лексики современного языка", отражаемый в средних словарях, хорошо соответствует объединению всех активных индивидуальных словарей. Наконец, большой словарь, "представляющий современный литературный язык в широкой исторической перспективе" хорошо соответствует объединению всех активных (продуктивных) и пассивных (рецептивных) индивидуальных словарей. В совокупность пассивных частей индивидуальных словарей попадают те единицы из уста-

ревшего лексического фонда, которые могут быть известны широкому кругу читателей по популярным литературным произведениям.

3. Каждый из трех основных типов толковых словарей отличается от других типов по объему словника, по семантическому объему слов, повторяющихся в них, и по стилистическим характеристикам слов. Но никто еще систематически не рассматривал вопрос о том, насколько последовательно в существующих словарях осуществляется принцип включаемости меньшего по объему словаря в больший, за счет каких словообразовательных категорий лексики разрастается объем большого словаря, с какими функционально-стилистическими сферами связана эта добавляющаяся лексика для среднего словаря в сравнении с кратким, большого в сравнении со средним и большого – с кратким. Ведь, как показывает опыт, в более краткий словарь иногда включается та часть лексики, которая может отсутствовать в большем по объему словаре, иногда в краткий словарь включаются такие значения у слов, общих с большим по объему словарем, которые не представлены в нем и т.п. Насколько систематически эти "искажения" проявляются, никто пока сказать не может.

При рассмотрении соотношения между словарями разного типа возникает и та проблема, что одна и та же информация в словарях разного типа может быть по-разному представлена. Основной задачей, которую мы ставим в данной работе, и является попытка проследить на определенном алфавитном участке двух словарей детальные соотношения по функционально-стилистическим и объемно-семантическим характеристикам между словарным составом и планом содержания слов, повторяющихся в двух словарях.

При трехступенчатой типологии толковых словарей выбор для сопоставления таких рядом стоящих типов, как средний и краткий, более предпочтителен, чем среднего и большого или, тем более, краткого и большого. В состав среднего и краткого толковых словарей входят слова, отражающие более центральную, более вычлененную из совокупного объема часть реального языкового сознания носителей данного языка, что в большей степени способствует объективности начального этапа подобного исследования.

Из указанных выше словарей были сделаны 4 выборки, включающие слова следующих алфавитных диапазонов: 'а' – 'баянист'; 'заветренный' – 'закаливать'; 'накрыться – 'нация';

'получиться' - 'похвальность'. В ходе исследования было проанализировано 3971 лексических единиц "Словаря русского языка" С.И. Ожегова (далее - СО), т.е. примерно по 1000 соответствующих единиц каждого из указанных алфавитных диапазонов и 6827 соответствующих по алфавиту лексических единиц из "Словаря под редакцией А.П. Евгеньевой (далее - МАС). Общее количество значений для слов данной выборки из СО составило 5580, для слов МАС'а - 9823.

Для сравнения данных словарей учитывались такие характеристики, как частеречная принадлежность слова, количество значений у слова в СО и в МАС'е, соответствия между значениями, оттенками значений и частями толкования слов, стилистические характеристики слова и его отдельных значений. Учет соответствия между значениями одного и того же слова в рассматриваемых словарях позволил выявить и систематизировать случаи расхождений (явных и неявных) в представлении семантического объема одного и того же слова. Обработка стилистических данных проводилась как для всей лексики в целом, (на данных алфавитных диапазонах), так и для отдельных групп слов, разделенных по частям речи. Особо учитывались случаи различий подачи омонимов в словарях, полисемические характеристики омонимичных слов.

4. Рассмотрим полисемические характеристики для всех слов данных алфавитных диапазонов в СО и МАС'е.

Как и ожидалось, и в СО, и в МАС'е большинство слов являются однозначными. В СО в данной выборке из 3971 слова содержится 2906 однозначных слов, а в МАС'е из тех же 3971 повторяющихся слов 2404 однозначных. Уменьшение числа однозначных слов в МАС'е объясняется тем, что ряд однозначных слов СО (а именно 492) в МАС'е перешел в другие полисемические разряды (подробнее см. ниже).

Рассмотрим соотношение общих полисемических распределений (все части речи вместе) в СО и МАС'е (см. таблицу I) по 3971 повторяющемуся в обоих словарях слову.

Эти, как и другие экспериментальные данные, показывают, что семантический объем слова и количество слов с данным семантическим объемом (который определяется количеством значений у слова) в словаре статистически связаны: наибольшую долю в словаре составляют однозначные слова, затем следуют в порядке убывания частот слова с двумя, тремя и т.д. значениями. "Такой вид распределения, - считает Ю.А. Тулдава,- представляет собой, по-видимому, универсальную квантитативно-

Таблица I

Общее полисемическое распределение для выборки в 3971 слово СО и МАС (все части речи вместе)

Полисемия	Количество слов с данной полисемией		Общее кол-во значений	
	СО	МАС	СО	МАС
I	2906	2404	2906	2404
2	747	964	I494	I928
3	I86	324	558	972
4	76	II8	304	472
5	30	67	I50	335
6	I8	26	I08	I56
7	3	I0	2I	70
8	2	5	I6	40
9	2	4	I8	36
I0	-	3	-	30
II	-	I	-	II
I5	I	-	I5	-
I6	-	I	-	I6
В с е г о	3971	3927	5590	6470

системную характеристику полисемии естественных языков. По существу, здесь, как и во многих других случаях, проявляется известный принцип концентрации и рассеивания лингвистических единиц"(Тулдава, I987, с. I34).

Сопоставление двух словарей показывает, что на одном и том же лексическом материале при наличии общей в них тенденции этого рода обнаруживается сдвиг в распределении слов в сторону большего удельного веса в нем слов с бóльшим количеством значений в МАС'е в сравнении с СО.

Соотношение объемов групп слов с разным количеством значений в МАС'е и СО в избранном алфавитном диапазоне выравнивается, если мы берем данные не только по тем словам, которые представлены и в МАС'е, и в СО, но и по тем, которые в пределах указанного алфавитного диапазона появляются в МАС'е как новые в сравнении с СО (см. табл. 2). Слова в большем по объему словаре, как правило, увеличивают свой семантический объем, происходит их общая передвижка в более высокие полисемические разряды. Кроме того, появляется определенное количество новых слов, относительно уравновешивающих образовавшийся дефицит в области однозначности и малозначности.

Таблица 2

Соотношение объемов групп слов с разным количеством значений в МАС'е и СО

				Полисемия													Итого
				I	2	3	4	5	6	7	8	9	I0	II	I5	I6	
Кол-во слов с данн. полис.	Имя сущ.	СО	I	379	I02	27	7	4	3	I	I						524
			2	966	I89	36	I0	4	I	I							I207
		МАС	I	302	I23	4I	I8	I4	4	I	I						504
			2	869	249	65	I5	5	5	I		I	I	I			I2I7
	Имя прил.	СО	I	I50	30	6	I										I87
			2	200	46	6	3	4									259
		МАС	I	I2I	4I	I6	2	I	I								282
			2	I58	67	I4	6	2	I		I						249
	Глаг.	СО	I	898	3I0	95	47	I7	I3	I	I	2			I		I485
			2	I85	38	I2	8	I	I								245
		МАС	I	70I	380	I63	62	39	I3	8	3	3	I			I	I374
			2	I42	65	I4	I3	4	I								239
	Нареч.	СО	I	92	I7	2											III
			2	I9	5	I											25
		МАС	I	80	20	4	I	I									I06
			2	I8	I0	4											32

+ Для каждого словаря указаны данные отдельно для слов с приставками (I) и без приставок (2).

Таблица 3

Полисемические переходы "СО-МАС" в процентном выражении и абсолютных цифрах (все части речи вместе)

СО \ МАС	I	2	3	4	5	6	7	8	9	I0	II...I6
I	75,I 2I82	I6,I 444	I,2 34	0,33 9	0,I5 4						
2	II,7 9I	60,2 468	I6,3 I27	I,7 I3	0,89 7	0,I3 I					
3	0,54 I	I9,9 37	42,5 79	I8,3 34	9,7 I8	I,7 3				0,54 I	
4		5,2 4	I4,3 II	44,2 34	I5,6 I2	9,I 7	2,6 2	I,3 I			
5		3,I I	6,3 2	9,4 3	43,8 I4	I9,4 6		3,I I	3,I I		3,I I
6			I6,7 3	II,I 2	II,I 2	I6,7 3	27,8 5	5,6 I			
7							33,3 I		33,I I		33,I I
8								50 I		50 I	
...											
I5											I00 I

5. Рассмотрим более детально, в какие полисемические разряды переходят слова в МАС'е из определенных полисемических разрядов СО (см. табл. 3).

Внимательное рассмотрение таблицы № 3 показывает, что рассеяние слов в МАС'е по разным полисемическим зонам является тем большим, чем более полисемичной является данная группа слов по СО. Это проявляется в том, что однозначные слова (75,I %) стремятся быть однозначными и в МАС'е; двузначными по МАС'у стремится быть значительно меньше однозначных (I6,I %); незначительное число случаев характеризует переход однозначных по СО слов в трех-, четырех- и пятизначные в МАС'е. Двузначные слова, хотя преимущественно и остаются в двузначных по МАС'у (60,2 %), но в меньшем проценте случаев, чем в своей же зоне однозначные, примерно в равной степени двузначные слова становятся однозначными (II,7 %) и трехзначными (I6,3 %); незначительно рассеяние слов по остальным полисемическим зонам (четырех-, пяти- и шестизначных слов по МАС'у). Еще менее выражено предпочтение к переходу в определенную полисемическую зону в словаре МАС трех-, четы-

рех- и пятизначных слов из СО (42-44 %); однако предпочтение продолжает отдаваться переходу в ту же полисемическую зону по МАС'у, что и та, в которой они находились в СО. Однако ситуация меняется, начиная с шестизначных слов по СО: в МАС'е они преимущественно представлены не в своей полисемической зоне (см., например, шестизначные слова по СО, которые больше всего, а именно, 27,8 % представлены в МАС'е в зоне семизначных слов). Семи- и восьмизначные слова по СО вообще представлены такими переходами в МАС'е, которые не концентрируются в какой-либо полисемической зоне.

Для характеристики переходов важным является и то, что в МАС'е слова практически не имеют превышения полисемии в МАС'е над полисемией в СО более, чем на четыре единицы. Вместе с тем отметим, что довольно значителен процент перехода слов в МАС'е в менее полисемичные зоны в сравнении с СО.

Аналогичным образом может быть рассмотрена картина полисемических переходов для слов каждой части речи по отдельности. Рассмотрение этого вопроса остается за пределами данной статьи.

6. За счет каких же значений увеличивается семантический объем слова в МАС'е по сравнению с СО? 227 значений из 1018 учтенных таким образом значений (исключаются те значения, которые появились у слов, бывших в СО или в МАС'е омонимичными каким-либо другим и переставшими быть таковыми в другом сопоставляемом словаре) имеют помету "разговорное", что составляет 22 % от общего числа новых значений повторяющихся слов в МАС'е в сравнении с СО. 103 значения имеют помету "устаревшее", что составляет 10,1 % от общего числа этих новых значений в МАС'е. 48 значений относятся к стилистическому разряду "книжно-специальные", что составляет 4,7 % от общего числа указанных новых значений в МАС'е. 8 значений относятся к стилистическому разряду "областные", что составляет только 0,8 % от общего числа новых значений. В целом 37,6 % от общего числа новых значений слов, повторяющихся в МАС'е в сравнении с СО имеют стилистические пометы.

Если учесть, что в целом в МАС'е стилистически помечены 32,4 % значений слов, то из этого вытекает, что специфические для МАС'а (на фоне СО) значения, как и ожидалось, являются более стилистически маркированными, менее нейтральными, т.е. далее отстоящими от ядра лексической системы русского языка, чем в целом значения слов всего МАС'а.

7. Однако изменение (в основном, увеличение) количества значений в МАС'е может являться следствием изменения и в способах описания значений в одном словаре в сравнении с другим.

Увеличение числа значений у слов в МАС'е по сравнению с СО может быть результатом более дробного членения в словарной статье одной и той же семантической информации.

Так, например, к словам "айва" и "алыча" в СО и МАС'е даются следующие толкования:

Айва

СО	МАС
Южное дерево с твердыми ароматными плодами, похожими по форме на яблоко, а также плод этого дерева	I. Южное плодовое дерево сем. розоцветных 2. Плод этого дерева, похожий по форме на яблоко или грушу

Алыча

СО	МАС
Южное плодовое дерево, близкое к сливе, а также плод его	I. Южное плодовое дерево 2. Плод этого дерева.

В этих случаях МАС представляет семантическую информацию более дифференцированным, что является, видимо, и более оправданным.

Аналогично соотношение семантических структур лексико-семантической группы "растение, дерево или кустарник, имеющее плод". В нашу выборку вошли такие слова из этой ЛСГ, как: абрикос, айва, алыча, ананас, арахис, арбуз, банан, барбарис, артишок, баклажан, помидор (всего II слов).

Другой пример перехода от однозначности в СО к двузначности в МАС'е:

Альбом

СО	МАС
Тетрадь в переплете для стихов, рисунков, каких-н. коллекций и т.п., а также собрание рисунков, репродукций (в виде книги, папки).	I. Тетрадь или книга с чистыми листами для стихов, рисунков, фотографий, открыток и т.п. 2. Объединенное по теме собрание иллюстраций, рисунков, репродукций, чертежей и т.п., переплетенных вместе или собранных в папку.

Здесь также увеличение числа значений в МАС'е в сравнении с СО не является следствием привлечения дополнительной семантической информации, а лишь отражает факт более корректного, более дифференцированного представления того же самого семантического содержания. Там и там содержится указание на целое (собрание рисунков, репродукций и т.п.) и на его часть (тетрадь в переплете для рисунков, репродукций и т.п.), но в СО они формально не разделены.

Из рассмотрения всего материала следует вывод, ЧТО СО довольно часто следует укоренившейся традиции более обобщенного представления семантической информации о словах, чем этого требует принцип отображения реального расслоения плана содержания слова на отдельные значения.+

Приведем также пример перехода от двузначности в СО к трехзначности в МАС'е, который является следствием такого же, как и выше, неоправданного "слияния" двух значений в одно в СО:

Атаман

СО	МАС
I. Название военно-административных должностей в казачьих областях до революции и в казачьих войсках в старину. 2. Главарь, предводитель.	I. В старину: выборный начальник казачьих дружин. 2. В дореволюционной России: назначаемый или выборный начальник в казачьих войсках и селениях, выполняющий военные, полицейские и административные функции. 3. Главарь, предводитель.

Нами был проведен анализ того, в каком числе случаев увеличение количества значений в МАС'е было произведено не за счет привлечения дополнительной (более периферийной) семантической информации, а за счет более корректного, более дифференцированного представления того же содержания, что и в СО. Таковых случаев оказалось 73. Кроме того в I2 случаях сверх этого мы сомневаемся в обоснованности "расщеплений" одного из "ожеговских" значений в МАС'е.

+ Обсуждение проблемы излишнего обобщения толкований слов в кратких словарях см. в (Шведова, 1981).

Можно констатировать, что хотя случаи "сверхобобщений" в СО и имеют место, но не они определяют решающим образом ситуацию с соотношением семантических объемов слов в сопоставляемых словарях, т.к. число новых значений в МАС'е, получаемых за счет указанного "расщепления" составляет всего чуть более 8 % от числа новых значений в нем у слов, имеющих соответствие в СО.

8. Особого внимания заслуживают случаи уменьшения семантического объема слова в МАС'е в сравнении с СО. Всего таких случаев 151. Анализ показывает, что уменшение полисемии в СО по сравнению с МАС'ом происходит по следующим причинам:

1) "слияние" двух или нескольких значений СО в одно значение в МАС'е (всего 18 раз);

2) выделение одного или нескольких прежних значений в оттенок значения в МАС'е (всего 67 раз);

3) различие в способах представления слов с фразеологически связанными значениями (1 раз);

4) комбинированные случаи (чаще всего (1) и (2)) (всего 5 раз);

5) неправомерная "ликвидация" некоторых специфических (или новых) значений в МАС'е в сравнении с СО (всего 60 раз);

а) новые значения, которые по неизвестным причинам не вошли в МАС; они могли и не войти в СО, но вошли (всего 2 раза);

б) значения производных слов, относительно нерегулярные, которые должны быть в МАС'е, но их там нет (всего 7 раз);

в) значения с пометами в СО, отсутствующие в МАС'е (всего 22 раза);

г) в МАС'е не отображено значение, устаревшее к моменту его переиздания, а в СО - отображено (2 случая);

д) случайные упущения в МАС'е в сравнении с СО (всего 33 раза).

Как видно, случаи уменьшения семантического объема слов в МАС'е немногочисленны, они составляют всего лишь 3,8 % числа слов, общих для обоих словарей. Это составляет 16,9 % всех случаев изменения семантического объема слов в МАС'е в сравнении с СО. Существенно отметить и то, что почти половину случаев уменьшения семантического объема в МАС'е в сравнении с СО связано с представлением самостоятельного

значения по СО в качестве оттенка значения в МАС'е. Т.е. реально удельный вес фактов, свидетельствующих о нетипичности для большого по типу словаря уменьшении семантического объема его слов, оказывается еще меньшим, чем это указано выше.

Появление новых значений в СО свидетельствует либо о регистрации неологизмов этим словарем, либо о том, что лексикографы обращаются к описанию периферии семантики слова, что нетипично для краткого словаря. Особенно наглядно это проявляется там, где новые значения у слов в СО имеют пометы "книжн.", "спец.", "устар.". Таковых по исследованному материалу оказалось, как указано выше, 22 слова. Кроме того 38 слов СО имеют новые значения, отсутствующие в МАС'е, без всяких помет.

9. В четырех выборках указанного алфавитного диапазона содержится 2900 слов, представленных в МАС'е, но отсутствующих в СО. 217 единиц из 2900 являются формами страдательного залога глаголов. В СО не даются формы страдательного залога на -ся, "так как они всегда образуются по правилам грамматики от глаголов несовершенного вида" (Ожегов, 1972, с. 11). В МАС'е глаголы страдательного залога даются отдельными словарными статьями, если они не омонимичны соответствующим возвратным глаголам. Всего, таким образом, в том алфавитном диапазоне, который рассматривался, на 3971 слово СО приходится 6827 им соответствующих слов МАС'а.

Большинство слов из 2900 однозначные. Однако однозначных среди новых слов существенно больше, чем среди тех, которые "пришли" из СО (67 % и 61 %). По количественной представленности слова разных частей речи распределяются так: имя прилагательное (1131, в т.ч. 135 приставочных), глагол (998, в т.ч. 909 приставочных), имя существительное (909, в т.ч. 293 приставочных), наречие (157, в т.ч. 82 приставочных), прочие части речи (9). Т.е. максимум полисемии у этих слов существенно ниже, чем у "перешедших" из СО. Ниже приводится таблица (№ 4) распределения полисемии для 2683 слов специфичных только для МАС'а (исключая формы страдательного залога глаголов; для слов всех частей речи вместе).

93 слова из 2900 помимо стилистически помеченных значений имеют также и нейтральные значения. Эти слова в большинстве представлены, во-первых, глаголами несовершенного вида, вторично образованными от приставочных глаголов совершенного вида, типа "закольцевать" (несов. к "закольце-

вать" при наличии в словаре "кольцевать"), во-вторых, отглагольными существительными, которые непоследовательно включались в СО. Теоретически эти слова в своих нейтральных значениях могли бы быть включены в СО.

Таблица 4

Полисемические распределения для 2683 слов, специфичных только для МАС'а

ПОЛИСЕМИЯ	КОЛИЧЕСТВО СЛОВ С ДАННОЙ ПОЛИСЕМИЕЙ
I	2214
2	881
3	81
4	13
5	11
6	3

10. В СО в данной выборке из 3971 слова содержится 115 слов, которые отсутствуют в МАС'е в противоречии с принципом полной включаемости словника краткого словаря в словник более полного. 60 слов из них имеют стилистическую пометы: 19-помету "просторечное", 22 - помету "разговорное", 3 - помету "специальное", 2 - помету "устарелое", 1 - помету "высокое", остальные - "уменьшительное". 20 слов относится к синтаксическим дериватам.

Бо́льшая часть этих слов, на наш взгляд, должна была бы быть представлена в МАС'е, т.к. относится к актуальной для современного языка лексике, а в ряде случаев - к числу ядерных слов ("заиметь", "заметано", "замешкать", "по-людски", "понятийный", "поруганный" и др.). Вместе с тем, в этом списке есть и некоторое количество слов, которые уже вышли из употребления и попали в СО, видимо, случайно (а точнее, по традиции, заимствуясь из основного источника словника для СО - из "Толкового словаря русского языка" Д.Н. Ушакова). Например: "народоведение", "народоведческий", "наливочный" и некоторые др. Из числа слов, которые имеют помету "разг.", но на самом деле являются полудиалектными, полупросторечными, т.е. находящимися на грани литературного языка, нами в рассматриваемой группе специфических только для СО слов обнаружено только два слова - "нарочно" и "поставный".

Эти факты свидетельствуют о том, что состав СО, видимо, в целом хорошо соответствует задачам, ставящимся перед сло-

16*

варем краткого типа, задачам отображения ядерного состава литературного словаря. Отклонения от выполнения им этой задачи являются мнимыми, связанными с особенностями представления информации в нем, либо случайными, неизбежно возникающими как вследствие отсутствия в реальности жестких границ между различными слоями литературного словаря, так и вследствие индивидуальных колебаний составителей словарей при реализации принципов их комплектации и описания их семантики.

ЛИТЕРАТУРА

Апажев М.Л. Лексикография и классификация словарей русского языка. Нальчик, 1971.

Денисов П.Н. Очерки по русской лексикологии и учебной лексикографии. М., 1974.

Копецкий Л.В. Теоретические предпосылки двуязычной славянской лексикографии. Praha, 1972.

Новиков Л.А. Типология учебных словарей // Учебные словари русского языка. II Международный конгресс МАПРЯЛ. Доклады советской делегации. М., 1973.

Ожегов С.И. О трех типах толковых словарей русского языка// Ожегов С.И. Лексикология, лексикография, культура речи. М., 1974.

Ожегов С.И. Словарь русского языка / Под ред. Н.Ю. Шведовой. 9-е изд. М., 1972.

Поликарпов А.А. Полисемия: системно-квантитативные аспекты // Квантитативная лингвистика и автоматический анализ текстов. (Ученые записки Тартуского гос. ун-та. Вып. 774). Тарту, 1987.

Поликарпова А.О., Поликарпов А.А. Опыт изучения и характера знания русской лексики // Квантитативные аспекты системной организации текста. Материалы межвузовского семинара. Тбилиси, 1986.

Шведова Н.Ю. Однотомный толковый словарь // Русский язык. Проблемы художественной речи. Лексикология и лексикография. М., 1981.

Щерба Л.В. Опыт общей теории лексикографии, этюд I // Щерба Л.В. Избранные работы по языкознанию и фонетике. Вып. I. Л., 1958.

Malkiel J. Distinctive Features in Lexicography: A Typological Approach to Dictionaries Exemplified with Spanish (II). Romance Philology, vol. XIII, No 2, Nov. 1959.

Rey-Debove J. Etude linguistique et sémiotique des dictionnaires français contemporains. The Hague-Paris: Mouton, 1971.

Sebeok T.A. Materials for a Typology of Dictionaries // Lingua, 1962, vol. XI.

ABOUT THE SYSTEMATIC RELATIONSHIP BETWEEN SMALL AND MEDIUM-SIZED EXPLANATORY DICTIONARIES OF RUSSIAN

A.A. Polikarpov, O.S. Kryukova

S u m m a r y

A typology of explanatory (monolingual) dictionaries is specified such as would reflect the functional-communicative stratification of the total vocabulary of a language into the following three layers: (1) "kernel vocabulary" made up by the shared part of the individual active normative vocabularies of all normal native speakers; (2) "the total vocabulary of a period of time", i.e. the sum total of these individual active vocabularies; (3) "the cumulative vocabulary of a people", made up by all the active and passive vocabularies of the normal native speakers of the corresponding language.

The issues under consideration include the relation between dictionary sizes (as measured in entry words), the semantic volume of words common to different dictionaries, etc. A tendency is noted for words to shift to a new polysemantic field when presented in a medium-sized dictionary as opposed to a small one.

ЗАГОЛОВОК И ЕГО ТЕМА-РЕМАТИЧЕСКОЕ ЧЛЕНЕНИЕ

В.А. Чижаковский

Заголовок научно-технических и научных текстов[+] может рассматриваться как особый вид предложения, представляющий собой отдельный компонент связного текста. Используя минимальное количество лексических средств, он прогнозирует информацию, которая заложена автором в текст статьи. Во многих случаях заголовок определяет тему и рему статьи (Чижаковский В.А., 1986, с. 65-108; Чижаковский В.А., 1988,с.11-13). Эти функции заголовка обеспечиваются следующими морфолого-структурными и семантико-синтаксическими особенностями:

1) номинативным (конкретным и абстрактным) характером большинства слов, входящих в заголовок;

2) структурной связью и взаимозависимостью лексических единиц, что обеспечивается предложенным и присубстантивным управлением и примыканием;

3) совпадением границ структурных и семантических сегментов заголовка;

4) зависимостью между позициями различных заголовочных компонентов и их функцией;

5) наличие лексических средств, по которым определяются границы сегментирования заголовка.

Кроме того, используется такая особенность заголовка как равномерное соотношение в сообщении известной обоим собеседникам информации (темы) и новых сведений (ремы), призванное обеспечить эффективность коммуникативного процесса. Тема-рематический анализ заголовка осуществляется на базе тезаурусных и тезаурусно-фреймовых моделей, служащих для выделения тематических индикаторов (ТИ) и списков абстрактных существительных или их рематических атрибутов, когда речь идет о сложных словах или словосочетаниях. Такие списки используются для выделения ремы (рематических индикаторов - РИ). Основное значение тема-рематического анализа заголовка заключается в том, что он в большинстве случаев по-

---

[+] Материалом исследования послужил 1741 заголовок немецкого специального журнала "Водное хозяйство и гидротехника" (**Wasserwirtschaft und Wassertechnik** - WWT) за период с 1976 по 1985 гг. и 671 заголовок научных статей по специальности "Инженерная лингвистика" и опубликованные за период с 1957 по 1985 гг. членами всесоюзной группы "Статистика речи" (Чижаковский В.А., Бектаев К.Б., 1986, с. 22-70).

зволяет формально ограничить тему от ремы путем установления отличительных семантических признаков, их выражающих лексических единиц. Если обнаруживается противопоставление конкретного значения одних существительных (тема) абстрактному значению других (рема), то это позволяет считать, что заголовки предсказывают в общих чертах "сюжет" статей. Например, в русском заголовке "Организация[I] словарной статьи[2] в многоуровневой системе[3] машинного перевода[4]"+ и немецком, опубликованным в журнале WWT № 2 за 1985 г. на с. 13-15, Entwicklung[1] des Wasserbedarfs[2] der Bevölkerung[3] bis zum Jahr 2010[4] (Развитие потребности в воде населения до 2010 года) противопоставляются абстрактные значения словоформ организация и Entwicklung (развитие), называющих ремы статей (они подчеркнуты одной непрерывной линией) конкретным значениям словосочетания словарной статьи и словоформы Wasserbedarfs (потребность в воде), выражающей темы статей (они подчеркнуты двумя непрерывными линиями). Такая ситуация позволяет реципиенту предугадывать примерное содержание соответствующих статей.

Состав структур исследуемых заголовков представляет собой поддающиеся автоматическому сегментированию отрезки, в число которых входят тематические и рематические компоненты. Эти отрезки были условно названы позициями и пронумерованы в соответствии с порядком их расположения (см. вышеприведенные примеры++).

Как в немецких, так и в русских исследованных заголовках предпочтение отдается первопозиционному расположению рематического компонента. Он предшествует тематическому компоненту (ТИ) в 1608 случаях из 1741, что составляет 92,36 ± ± 1,23 %+++ для немецких заголовков и в 581 случае из 671,

---

+ Аполлонская Т.А. Организация словарной статьи в многоуровневой системе машинного перевода // Инженерная лингвистика и преподавание иностранных языков с помощью ТСО. - Л., 1981. - С. 66-83.

++ Сегментирование заголовка происходит на основе автоматического выделения по формальным признакам границ отрезков. Кодированная информация о наличии таких признаков вводится в память ЭВМ. Цифры 1, 2, 3, 4 обозначают эти границы и свидетельствуют о том, что в каждом из названных заголовков имеются по четыре позиции.

+++ Абсолютная ошибка определяется по формуле: $\varepsilon = Z_p \sqrt{\frac{f(1-f)}{N}} \cdot 100$, где ε - абсолютная ошибка; Z_p - константная величина; она равна 1,96 при p (надежность) равное 0,95; f - относительная частота; N - количество исследуемых заголовков (Piotrowski R. u.a., 1985, s. 404-410).

что составляет 86,58 ± 2,74 % для русских заголовков (см. таблицу I).

Такое рема-тематическое структурное построение заголовка можно объяснить стремлением автора стимулировать интерес читателя к содержанию документа с первого же его предъявления. Поскольку такая "притягательная сила" (Вайнрих Х.,1978) выражается ремой заголовка, то ее первопозиционному месту расположения отдается предпочтение (см. вышеприведенные статистические данные). Что касается подбора соответствующих коммуникативному назначению рематического компонента лексических средств (абстрактных существительных, атрибутивно-рематических прилагательных или других знаменательных слов, предлогов или слитных артиклей и знака тире), то этот вопрос требует более подробного рассмотрения. Ему посвящен материал настоящей статьи.

Известно, что индивидуальное восприятие новой информации неодинаково у разных читателей (Тоом А.Н., 1976, с. 112-126; Гончаренко В.В., Шингарева Е.А., 1984, с. 53-55). Поэтому автор статьи стремится быть носителем того общего, что объединяет его с основной массой читателей-специалистов и обеспечивает контакт для передачи своего замысла на основе новых идей, которые он должен выразить словесно, пользуясь общепринятым и общедоступным материалом. Пространственная ограниченность заголовка позволяет ему определить только лингвистические "ориентиры" поиска информации в сопровождаемом тексте.

Как показало исследование, для их номинации и выделения, он обращается к таким лексическим средствам как абстрактные существительные, атрибутивно-рематические прилагательные или другие знаменательные слова, предлоги или слитные артикли, имеющие предложно-рематическое значение и знак тире, за которым следует рематический компонент.

I. Абстрактные существительные как словесные выражения процесса абстракции, предполагающего мысленное отвлечение от несущественных свойств какого-либо предмета и выделение его основных связей и отношений, позволяют проявление различия и разнообразия в толковании последних. Данное семантическое свойство абстрактных существительных используется автором заголовка в количестве РИ. Такие имена существительные как анализ, опыт, проблема, развитие и т.д. (**Analyse, Erfahrungen, Problem, Entwicklung**) отражают своей отвлеченностью возможность разнообразного подхода к толкованию свойств на-

званных сущностей. Известно, что чем больше в какой-либо совокупности различных элементов, тем более в ней содержится информации (Урсул А.Д., 1975). В русском заголовке Опыт машинного перевода английских и японских научных текстов (Чижаковский В.А., 1986, с. 30) и немецком **Erfahrungen beim Einsatz von Pflanzenbecken**[+] (Опыт использования зарослих водоемов) в качестве первопозиционных РИ абстрактные существительные опыт и **Erfahrungen** (опыт). Понятие 'опыт' включает какую-то "совокупность знаний и общественной практики людей" и "является основой познания и критерием ценности наших знаний об окружающем мире" (Кондаков Н.И., 1975, с. 413).Таким образом, оно само по себе допускает разнообразие в познании истинности и тем самым возбуждает наш интерес вообще. Если же его сочетать с такими конкретными понятиями как 'машинный перевод' (см. первый заголовок) и 'использование зарослих водоемов' (см. второй заголовок), то становится ясным намерение автора посредством слова опыт раскрыть сущность содержания новизны сообщения и сориентировать читателя-специалиста в поисках данной новизны в самом тексте.

Сопоставительные статистические данные позволяют говорить о более частотном использовании для выражения РИ абстрактных существительных в немецком языке по сравнению с русским: 58,30 ± 1,08 % и 25,05 ± 2,76 % случаев (см. табл. 1, пункт 1, 2, 6а и 6б).

2. Имена прилагательные и другие знаменательные слова, сочетающиеся с именами существительными (обычно с конкретным значением) или входящие в состав сложных немецких слов, где определяемое слово имеет, как правило, тоже конкретное значение, выполняющее чаще всего атрибутивно-рематическую функцию. Они характеризуют сопровождаемое существительное признаком, значение которого, по мнению автора, должно привлечь внимание читателя к рассматриваемой в статье теме своей необычностью, как, например, в словосочетаниях **verbesserte Algorithmen** (улучшенные алгоритмы) и частотные словари прилагательные **verbesserte** и частотные. Названные прилагательные характеризуют сопровождаемые существительные такими свойствами, которые с одной стороны, придают последним качества, в котрых могут быть заинтересованы читатели-специалисты, с другой - вносят разнообразие в их оценки.

[+] См. журнал **WWT**, № 4, 1985, с. 82.

В тех случаях, когда определяемое существительное имеет абстрактное значение, то определяющее слово выполняет обычно атрибутивно-тематическую функцию, как, например, в словосочетаниях **aerobe Stabilisierung** (аэробная стабилизация) и автоматический валентностный анализ, то есть позволяет абстрактным существительным участвовать в формировании ТИ (подробнее см. в Тоом А.Н., 1976; Урсул А.Д., 1975).

Установленные выше соотношения значений между определяемым и определяющими словами не носит абсолютный характер. Если взять, например, немецкий рематический компонент **neue Beispiele** (новые примеры) и русский предварительные итоги, то в обоих случаях значение элементов РИ рематическое (атрибутивно-рематическое и абстрактное).

Общие статистические данные употребительности РИ включающих элементы с атрибутивно-рематическим и атрибутивно-тематическим значениями, свидетельствуют об их меньшей частотности в немецких заголовках по сравнению с русскими: 20,30 ± 1,86 % и 44,35 ± 3,92 % случаев (см. табл. I, пункты 3, 4, 5, 6в, 6г).

Сочетание атрибутивно-рематического элемента с определяемым абстрактным существительным позволяет передать первопозиционному компоненту заголовка дополнительное тематическое значение. Таким образом, в одном и том же компоненте совмещаются рематическое и тематическое значения. Последнее, т.е. тематическое значение, присовокупляется в том случае, когда хотя бы один из элементов первопозиционного компонента входит в тезаурусную или тезаурусно-фреймовую модели той или иной предметной области (подобласти), определяющие их тематическую принадлежность. В противном случае он остается только рематическим индикатором. Например,в немецком заголовке **Ökonomische Bewertung der Wasserressourcen** (Экономическая оценка водных ресурсов)[+]. начальный компонент **Ökonomische Bewertung** представляет собой словосочетание абстрактного существительного **Bewertung** (оценка) с атрибутивно-тематическим прилагательным **ökonomische** (экономический). Такое словосочетание входит в тезаурусную модель предметной подобласти "Экономика водного хозяйства" журнала **WWT**[++].

[+] См. журнал **WWT**, № 7, 1985, с. 149-150.

[++] Семантическое пространство журнала **WWT** представлено предметной областью "Водное хозяйство и гидротехника" и включает девять предметных подобластей, одна из которых является "Экономика водного хозяйства".

Наличие тематического значения в РИ не исключает использование в заголовке и другого тематического индикатора, как это имеет место в данном примере. Словоформа **Wasserressourcen** (водные ресурсы) входит в тезаурусную модель предметной области "Водоснабжение", которая также является одной из составных частей предметной области, представляющей журнал **WWT** (см. вторую сноску).

Использование в заголовке двух ТИ следует рассматривать как отражение комплексности описываемой в статье тематики. Она включает не только вопросы экономики водного хозяйства вообще, но и конкретные экономические оценки наличных водных ресурсов.

3. Как можно прочитать в табл. I (см. пункты 7, 8) первопозиционный РИ может начинаться с предлога или слитного артикля, а также следовать после знака тире.

В первом случае (см. пункт 7), автор подчеркивает таким способом свою точку зрения и отношение к обсуждаемому вопросу, что должно напомнить читателю о возможно новом подходе к рассматриваемой теме. Употребительность такого способа построения первопозиционного РИ составляет 5,I0 ± I,08 % или 89 случаев в немецком заголовке и соответственно I7,I0 ± 2,84 % или II4 случаев в русском. Такие предлоги или слитные артикли характеризуются предложно-рематическим значением.[+]

Во втором случае речь идет о той позиции заголовка, которая стоит после знака тире и условно названа первопозиционной. Роль знака тире заключается в том, чтобы служить формальным признакам выделения РИ заголовка, который в таких случаях всегда занимает позицию после данного знака. Последний характеризуется рема-сигнальным значением. Статистические данные свидетельствуют о более частотном употреблении такого способа ввода РИ в немецких заголовках по сравнению с русскими (см. пункт 8). Они составляют I54 случая, или 8,9 ± I,35 % в немецких заголовках по сравнению с тремя случаями, или 0,45 ± 0,5I % в русских.[++]

[+] **Zur Ausgrenzung von Trinkwassergebieten** (К вопросу об установлении границ водоохранных зон питьевой воды) (см. **WWT**, № 4, I985, с. 77-878); Об одном способе автоматического выявления лексической разнородности текста (Чижаковский В.А., Бектаев К.Б., I986, с. 5I).

[++] **Wasser - eine kostbare Gabe der Natur** (Вода - ценный подарок природы) (см. **WWT**, № 3, I985, с. 50-5I); Терминологический взаимообмен - закономерный процесс) (там же, с. 6I).

Сравнительные статистические данные об употребительности первопозиционного рематического индикатора исследованных немецких и русских заголовков позволяет говорить об их предпочтительном использовании в таких случаях в обоих языках. Рема-тематическое структурное построение вышеназванных заголовков и возможность распознавания их компонентов способствует формальному выделению соответствующих структурных единиц (позиций), которое может быть реализовано на ЭВМ.

Л И Т Е Р А Т У Р А

Вайнрих Х. Текстовая функция французского артикля / Пер. с анг. // Новое в зарубежной лингвистике. - М., 1978. - Вып. VIII. - С. 370-387.

Гончаренко В.В., Шингарева Е.А. Фреймы для распознавания смысла текста. - Кишинев: Штиинца, 1984. - 198 с.

Кондаков Н.И. Логический словарь-справочник. - М.: Наука, 1975. - 720 с.

Тоом А.Н. Несимметричная коммуникация, формализация и управление в играх // Семиотика и информация. - М., 1976. - Вып. 7. - С. 112-126.

Урсул А.Д. Проблема информации в современной науке. Философские очерки. - М.: Наука, 1975. - 288 с.

Чижаковский В.А. Заголовок - составной элемент текста в коммуникативной системе "человек-машина-человек". - Кишинев, 1987. - 156 с. Рукопись представлена Кишиневским сельскохозяйственным институтом им. М.В. Фрунзе и депонирована во ВИНИТИ 29 октября 1986 г., № 720,М.

Чижаковский В.А. Семантико-коммуникативные аспекты автоматической переработки заголовка научно-технического текста: Автореф. дис. д-ра филол. наук. - Л., 1988. - 31 с.

Чижаковский В.А., Бектаев К.Б. Статистика речи 1957-1985 гг.: Библиографический указатель. - Кишинев: Штиинца, 1986. - 111 с.

Piotrowski R., Bektaev K., Piotrowskaja A. Mathematische Linguistik. - Bochum: Studienverlag Dr. N. Brockmeyer, 1985. - 514 S.

Таблица I

Первопозиционный рематический компонент заголовков научно-технических и научных немецких и русских текстов (их морфолого-синтаксический и семантический состав)

№/№ пп	Формы выражения и их значения	Примеры и количественные показатели					
		немецкий язык		%	русский язык		%
I	2	3	4		5	6	7
I.	Простое слово		557	32,00	сопоставление	88	I3,00
	существительное абстрактное	**Möglichkeit** (возможность)	96	5,50	автоматизация	74	II,00
		Reinigung (очистка)					
2.	Сложное слово						
	существительное+существительное я (определяющее) (определяемое)						
а)	абстрактное абстрактное	**Nutzungsmöglichkeit** (возможность использования)	I7,2	9,90	нет	-	-
б)	конкретное абстрактное	**Chlorbestimmung** (определение содержания хлора)	66	3,80	нет	-	-
в)	абстрактное конкретное	**Ausführungstechnologie** (производственная технология)	I7	I,00	нет	-	-

Продолжение таблицы I

I	2	3	4	5	6	7
3.	Сложное слово знаменательное слово+существительное (кроме существительного) (определяющее) (определяемое)	**Wiederverwendung**				
а)	атрибутивно-рематическое абстрактное	(повторное использование)	7	0,40	нет -	-
б)	атрибутивно-рематическое конкретное	**Messeinrichtung** (измерительное устройство)	I9	I,I0	нет -	-
4.	Сложное слово приложение					
а)	предложение предшествует основному слову	**KDT-Empfehlungen**				
	атрибутивно-рематическое абстрактное (или конкретное)	(рекомендации КДТ)	7	0,40	школа-семинар 3	0,45
б)	Приложение следует за основным словом	**Prozessanalyse Wasserverteilung** (процессуальный анализ водораспределения)	7	0,40	Подсистема "Абитуриент" 2	0,30
	абстрактное атрибутивно-тематическое					
5.	Словосочетание без союза (и) прилагательное+существительное (определяющее) (определяемое)					
а)	атрибутивно-рематическое абстрактное	**Neue Beispiele** (новые примеры)	II3	6,50	Предварительные итоги I3	I,95

Продолжение таблицы I

I	2		3	4		5	6	7
б)	атрибутивно-тематическое	абстрактное	**Ökologische Modelle** (экологические модели)	64	3,70	Семантическая модель	I22	I7,90
в)	атрибутивно-рематическое	конкретное	**Verbesserte Algorithmen** (улучшенные алгоритмы)	62	3,60	Частотные словари	I4	2,50
г)	атрибутивно-тематическое	абстрактное	**Aerobe Stabilisierung** (аэробная стабилизация)	I6	0,90	Автоматическое индексирование Формальный валентностный анализ	3	0,45
	Неопределенное местоимение (+прилагательное)+существительное							
д)	атрибутивно-рематическое (атрибутивно-тематическое)	абстрактное	нет			Некоторые вопросы Некоторые статистические характеристики	2I	3,I5
6.	Словосочетания с союзом **und** (и) Существительное+союз+существительное							
а)	абстрактное	абстрактное	**Aufgaben und Ergebnisse** (задачи и результаты)	I05	6,I0	Применение и оценка	7,0	0,95
	Существительное+союз+прилагательное+существительное							

Продолжение таблицы I

I	2	3	4	5	6	7
б)	абстрактное атрибутивно-тематическое абстрактное	(анализ и математическое моделирование)	I7	I,00	нет	-	-
	Прилагательное+существительное+союз+(прилагательное)+существительное						
в)	атрибутивно-тематическое (рематическое) абстрактное (или конкретное)	**Wasserwirtschaftliche Ergebnisse und Aufgaben** (водохозяйственные результаты и задачи)	I2	0,70	Структурная организация и принципы	4	0,60
	Прилагательное+союз+прилагательное+(прилагательное)+существительное	**Bautechnische und bautechnologische Probleme**					
г)	атрибутивно-тематическое атрибутивно-тематическое (атрибутивно-рематическое) абстрактное абсолютное (или конкретное)	(строительно-технические и строительно-технологические проблемы)	24	I,40	Восходящий и нисходящий подходы Английский и русский частотные словари	8	I,20
7.	Предлоги или слитный артикль+существительное (или прилагательное +существительное; может быть союз+слитный артикль+существительное)	**Zur Anwendung**					
	Предложно-рематическое абстрактное (атрибутивно-рематическое или тематическое и абстрактное или конкретное; предложно-рематическое и абстрактное)	(к вопросу о применении) **Zum Stand und zur Entwicklung** (к вопросу об уровне и развитии)	89	5,I0	К вопросу Из опыта о тезаурусном моделировании	II4	I7,I0

Продолжение таблицы I

I	2	3	4	5	6	7
		Mit hohem Leistungszuwachs (с высоким приростом производительности)				
8.	Знак тире, за которым могут следовать артикль, прилагательное и существительное или партицип П					
а)	рема-сигнальное атрибутивно-рематическое абстрактное	**- ein effektives Verfahren** (эффектный метод)	I54 8,90	Терминологический взаимообмен - закономерный процесс	3	0,45
б)	рема-сигнальное атрибутивно-рематическое конкретное	**Hauptinstrument** (главный инструмент)		нет	-	-
в)	рема-сигнальное абстрактное абстрактное	**- Möglichkeiten und Grenzen** (возможности и границы)		нет	-	-
г)	рема-сигнальное атрибутивно-рематическое	**- dargestellt** (представленный)		нет	-	-
		ИТОГО:	I608 92,60		58I	86,95

THE TITLE AND ITS RHEME AND THEME SEGMENTATION

Valentin A. Chizhakowsky

The subject-matter of the article is the communicative function of the title and the ways of its expression. The theme and rheme analysis of the title is realized on the basis of thesaurus and thesaurus-frame models which are used for theme-indicator building and lists including abstract nouns or concrete nouns accompanied by rhematic attributes used to mark out the rheme-indicators.

The realized research work permits us to state that the overwhelming majority of scientific articles titles includes both theme - and rheme - indicators and that their location there is strongly fixed. As a rule (for German and Russian languages) the rheme-indicator precedes the theme-one.

The possibility to formalize their detection allows us to use computers when solving such problems.

The article includes statistical data which confirm the revealed regularities.

Х Р О Н И К А

СТАТИСТИЧЕСКАЯ ЛЕКСИКОГРАФИЯ И УЧЕБНЫЙ ПРОЦЕСС

Н.Г. Милых

С 28 по 31 января 1989 г. в Киеве при КГПИИЯ состоялся научно-методический семинар на тему "Статистическая лексикография и учебный процесс", в котором приняли участие в основном члены коллектива составителей Лексико-грамматического частотного словаря английского языка (глагол). Коллектив уже 20 лет работает на основе договоров о творческом содружестве и включает 48 кафедр из 7 республик страны. В результате статистического обследования выборки в 10 млн словоупотреблений создана картотека, на базе которой составлено пять пробных тетрадей Частотного словаря сочетаемости английского языка, издан "Справочник наиболее употребительных английских словосочетаний" ("Просвещение", 1986), включающий 426 самых частых слов. Сейчас составляется частотный словарь словоизменительных форм английского глагола. Планируется создание серии частотных словарей сочетаемости для учебных и научных целей.

В семинаре принимали участие также проф. Ю.А. Тулдава (Тарту), проф. А.В. Зубов (Минск) и проф. С.Д. Береснев (Одесса), работающие в области лингвистической статистики и лексикографии.

Вступительное пленарное заседание открыла проректор по научной работе проф. М.П. Дворжецкая. С первым докладом "Проблемы методологии современной квантитативной лингвистики" выступил проф. Ю.А. Тулдава, который остановился на методологии, математических теориях и методах исследования. Докладчик выделил уровни методологии по степени обобщенности: философский, общенаучный, конкретно-научный, методики и техники исследования. Из проблем он выделил системность, синергетику, интерпретацию и междисциплинарность, особенно на уровне текста.

Доклад проф. В.А. Зубова был посвящен учебным частотным словарям в системе компьютерного обучения. Отметив ограниченность возможностей машины, докладчик указал на три возможных подхода к работе с ЭВМ: узко ориентированные словари и грамматики, использование правил формальной и математической логики, создание лингвистических баз знаний. Он обратил внимание на то, что нужны новые типы словарей, особенно для общественных наук.

В докладе В.И. Перебейнос, З.П. Рукиной, С.С. Хидекель "Типы учебных словарей" отмечено, что частотные словари (ЧС) имеют большую дифференцирующую, но не обобщающую силу, поэтому надо создавать серию ЧС-справочников для разных специальностей учащихся: ЧС слов, грамматических форм слова, словообразовательных моделей, моделей словосочетаний и конкретных словосочетаний.

На семинаре работало четыре секции. В секции № I "Использование частотных словарей в процессе обучения иностранным языкам" доклады строились вокруг "Справочника наиболее употребительных английских словосочетаний". И.М. Слонимская (Пятигорск) отметила, что справочник используется недостаточно широко в силу своей психологической непривычности, нужны методические рекомендации по работе с ним. Т.А. Бунтина (Донецк) подчеркнула, что словарь этот дает основу для любого текста. А.С. Хуршудянц (Ставрополь), Н.В. Войко, Т.В. Тихонова, А.П. Мусиенко (Киев), Е.Н. Дубнина (Москва) показали возможности использования справочника в учебном процессе и в научной работе студентов. В докладе С.С. Хидекель (Москва) говорилось о технике и программе составления учебных словарей, об их недостатках и задачах, о необходимости типизации учебных словарей.

На секции № 2 были заслушаны доклады М.Р. Кауль (Москва) "Глубина и типы именных словосочетаний высокочастотных существительных", Л.Р. Вайнера (Могилев) "Устойчивость словосочетания и частотность употребления", А.С. Хуршудянц "Сочетаемость частотных слов как критерий выявления системных связей в лексике", В.И. Мамедовой (Ростов-на-Дону) "О некоторых особенностях выделения моделей глагольной сочетаемости в различных функциональных стилях", М.М. Мамаенко (Киев) о словаре подъязыков геохимии и физики минералов, Н.В. Войко о юридической терминологии в учебных словарях. Н.В. Войко пришла к выводу, что юридические учебные словари-минимумы не отражают действительной частотности терминологических словосочетаний и неудачно их подают, что лишает словари учебной ценности. Секция приняла решение проинформировать издательства, выпускающие учебные словари, о необходимости более тщательной оценки и проверки словарей перед публикацией.

На секции № 3 "Статистическая лексикография и стиль" были отражены два направления в статистических исследованиях: I) статистическая оценка характеристик различных подъязыков: С.И. Мартынова (Ростов-на-Дону); 2) изучение функ-

ционирования отдельного слова: А.В. Ширикова (Ростов-на-Дону), В.П. Мазур (Херсон), или группы слов: Т.А Бунтина (Донецк), И.В. Тименко (Киев), Т.А. Мизин (Кировоград) в различных стилях и определение статистических параметров этих стилей на основе разноуровневых характеристик изучаемых лексем. Л.С. Рудакова (Орел) анализировала частотность морфологических форм глагола в зависимости от стиля, а в пределах одного стиля - от семантики глагола.

Секция № 4 была посвящена системным и функциональным характеристикам грамматических форм и категорий. В полемическом докладе "Еще раз о послелогах и фразовых глаголах" Н.Г. Милых (Ростов-на-Дону) говорила, что фразовый глагол - такое же фразеологическое образование, как и любое другое, вопрос его выделения спорен. Доклады Е.И. Гороть (Луцк) и М.В. Корнышевой (Москва) были посвящены анализу грамматических форм существительных, зависимости между семантикой и формой. Т.Е. Шевченко (Киев) исследовала зависимость между частотностью глагола и полнотой реализации глагольной парадигмы. Авторы доклада "Исчисление словоизменительной парадигмы английского глагола" В.И. Перебейнос и И.С. Кесельман предложили набор из II дифференциальных признаков (ДП): залог, время, наличие отрицательной частицы и др. На их основе исчислено более 4000 словоизменительных форм глагола и их регулярных вариантов, из которых актуализируется примерно 10 %.

Заключительное пленарное заседание было посвящено в основном теории и практике лексикографии. Проф. С.Д. Береснев говорил о семантических взаимоотношениях компонентов словосочетаний из двух и более существительных в научном тексте. Л.А. Тицкая (Донецк) поделилась опытом применения ЭВМ в статистической лексикографии. И.М. Слонимская отметила, что традиционные виды словарей перестают удовлетворять потребности лингвистов и изучающих языки. Необходимы словари нового типа, основанные на статистической лексикографии, более полно описывающие слово, его значение, лексико-грамматические, словообразовательные, стилистические и другие свойства.

Под руководством проф. Ю.А. Тулдава был проведен круглый стол на тему "Будущее статистической лексикографии".

STATISTICAL LEXICOGRAPHY AND TEACHING PRACTICE

V.G. Milykh

S u m m a r y

A scientific methodological seminar "Statistical lexicography in connection with teaching practice" was held in Kiev on January 28-31, 1989. Plenary meetings were devoted to the general research problems while four sections dealt with 1) the use of frequency dictionaries in language teaching; 2) statistic research in combinability; 3) statistical lexicography and style; 4) systemic and functional characteristics of grammatical forms and categories. A Round Table discussion headed by Prof. J. Tuldava was also held on the perspectives of statistical lexicology and lexicography.

РЕЦЕНЗИИ

ПРОБЛЕМЫ КВАНТИТАТИВНО-СИСТЕМНОЙ ЛИНГВИСТИКИ.

Рецензия на книгу: Ю.А. Тулдава. Проблемы и методы квантитативно-системного исследования лексики. – Таллинн: Валгус, 1987. – 204 с.

Рецензируемая монография представляет итоги и проблемы квантитативно-системных исследований лексики, предпринятых автором диссертации на протяжении последних 15 лет. В значительной степени это и итог развития квантитативной лингвистики последних десятилетий, ибо нет ни одной сколько-нибудь крупной проблемы этого раздела языкознания, которая бы ни была принята во внимание и не включена в теоретический и методологический контекст представленного исследования. Однако в работе кроме рационального и критического осмысления результатов квантитативного анализа некоторых областей организации лексики содержится и углубленная проработка ряда либо совершенно новых, либо малоисследованных областей: закономерности исторического роста и развития лексики, общая картина социальной дифференциации лексики по сферам употребления, закономерности соотношения численности различных функционально-стилистических помет слов в словаре литературного языка, зависимость роста объема словаря от роста текстовых массивов, закономерности распределения в словаре слов с различным количеством значений, закономерности распределения в словаре текста слов разной частоты, зависимости между частотой слов и их полисемией, частотой и длиной, частотой и словообразовательным потенциалом слова, лексико-стилистический анализ эстонских текстов, квантитативно-типологическое сопоставление текстовых и словарных характеристик эстонского языка с рядом других угро-финских и индо-европейских языков и т.д.

Это позволяет определить данную работу как оригинальное и глубокое исследование актуальных проблем современной лингвистической науки, представляющее новое, весьма перспективное направление в лингвистической науке, – квантитативно-системный подход. Этот подход не только осуществляется в данной работе, но и обосновывается. Синтез категорий системности и количественности в исследовании языка совершенно естественен и закономерен. Это знаменует восхождение языкознания на новый уровень развития. Квантификация анализа языка, как показывается в данной работе – это не некий "довесок" к

прежнему методическому аппарату, ориентированному преимущественно на "качественную" проблематику, а совершенно необходимый момент становления аппарата науки, изучающей системный объект (каковым и является язык). В любой системе количественные и качественные характеристики находятся в весьма органичном соответствии друг с другом. Познание количественных характеристик - это познание в специфически отраженном виде и качественных характеристик. Тонкие градации качественных свойств системного объекта могут быть в конечном счете познаны только через квантификацию интенсивности их проявления в объекте. Точный количественный анализ позволяет углубить представление о природе объекта, а выявленные закономерности в его организации позволяют делать весьма конструктивные и практически значимые прогнозы о возможных качественных состояниях объекта (в т.ч. и о его типологических перестройках) при наличии тех или иных учитываемых условий. Этой цели и служит развиваемый автором книги подход.

Системность этого подхода заключается в понимании языка как одного из тех объектов в природе, для которых характерным является целостность, тонкое и сложное взаимодействие и взаимная адаптация таких сторон их организации, как структура (сети связей, устойчивые взаимодействия), субстанция (элементы с их качественными характеристиками, входящие в объект-систему) и функционирование объекта, сложное взаимодействие его подсистем между собой и его как целого с внешней средой. Совершенно прав автор книги, когда он пишет, что (стр. IO) "проявление системного подхода в наши дни связано в первую очередь с необходимостью исследования больших сложных систем, которые, как правило, слабо структурированы и которые содержат частично неформализуемые элементы, причем функционирование таких систем часто происходит в условиях неопределенности. Именно речевая деятельность в целом, а также ее подсистемы, в том числе лексика, относится к таким сложным образованиям, познание которых настоятельно требует системного подхода".

Более точное определение для подобных систем - вероятностные. В работе подробно рассматриваются (и впоследствии на конкретном экспериментальном материале подтверждаются) основные типологические особенности этого рода систем, характеризующихся не только случайностью их параметров (низший уровень организации), но и определенной устойчивостью и регулярностью в массе случайных событий (высший уровень орга-

низации). В этих особенностях вероятностных систем проявляется диалектическая связь категорий случайности и необходимости. Предметом исследования лексики (или любой другой подсистемы языка) при квантитативно-системном подходе являются квантитативные свойства и закономерности строения и функционирования лексики, рассматриваемые с системных позиций и с упором на вероятностную природу функционирования языка. Лексика рассматривается как вероятностная система, в том числе со свойствами устойчивости и вариативности. Подобно всем системам лексика распадается на различные подсистемы и может рассматриваться как многоплановое и многоуровневое образование. В то же время она сама является подсистемой общей системы языка и составляет определенный уровень в иерархии языковых явлений.

Впоследствии, при анализе ряда подсистем и аспектов лексики и ее в целом это определение предмета рецензируемого исследования наполняется конкретным разносторонним содержанием. Это отличительное свойство данной работы, выполненной под единым перспективным углом зрения, по сложному, но единому плану с полным подтверждением заявленных принципов и постулатов, с открытием ряда перспективных специальных проблемных областей, которые становятся видимыми именно с указанных методологических позиций.

Одним из наиболее важных методологических достижений данной работы, как нам представляется, является критика узкого понимания системности языка. Когда нередко говорят о явлениях "антисистемности" в языке[+], а также о "разной степени системности языковых явлений"[++], то, как правило, оперируют фактами, вырванными из целостного диалектического контекста взаимодействия необходимости и случайности в системе, диалектического соотношения явления и сущности. То, что часто на первый взгляд выглядит "нерегулярным", "нерациональным", "противоречащим"[+++] порядку в системе, перестает казаться таковым, оказывается в совокупности с други-

---

[+] См. Будагов Р.А. Система и антисистема в науке о языке // ВЯ, 1978, № 4, с. 3-17; Филин Ф.П. Некоторые вопросы современного языкознания // ВЯ; 1979, № 4, с. 19-28.

[++] Общее языкознание. Внутренняя структура языка. - М.: Наука, 1972. - 565 с.

[+++] См. Маслов Ю.С. Введение в языкознание. - М.: Высшая школа, 1975. - 327 с.

ми фактами образующим закономерные вариативные ряды, если осознается сущность вероятностных систем с их устойчивостью основных параметров и тенденций и достаточной автономностью и вариативностью характеристик подчиненных параметров. В связи с этим Ю.А. Тулдава совершенно справедливо пользуется разграничением "ядра" и "периферии" в уровневой организации языка, относя к последней все нерегулярности и кажущиеся антисистемные явления (с. 13). Те, кто не осознают подобного принципа устройства вероятностных систем, образно выражаясь, за "деревьями" фактов нерегулярности периферийного устройства языка не видят "леса" всей системы, которая в наибольшей степени представлена "ядерными", центральными уровнями организации языка.

Так что пользоваться зауженным пониманием системности и потом сокрушаться, что язык ему не соответствует, что он "не вполне система" (так поступают выше цитированные авторы), неконструктивно в своей основе. Анализируемая диссертация объясняет эту ситуацию в лингвистике и почти все ставит на свои места. "Почти" потому, что, на наш взгляд, автором монографии все-таки не до конца прояснен вопрос о том, каков механизм возникновения и взаимодействия случайных и закономерных тенденций в поведении отдельных элементов системы, хотя все предпосылки для такого объяснения в рецензируемой работе представлены. Дело в том, что в вероятностных системах, будучи достаточно автономными, элементы испытывают как организующее влияние со стороны системного целого, так и влияние со стороны случайных, не имеющих отношения к системе факторов. В каждом элементе поэтому причудливым образом переплетаются необходимые, "навязанные" системой черты, и случайные, совершенно не предсказуемые. Однако правильный ракурс – рассматривать не каждый элемент в отдельности, а в совокупности, ансамблем – ведет к проявлению действительно замаскированных системных особенностей: при совместном "ансамблевом" рассмотрении элементов системно-закономерные черты в их организации будут повторяться, усиливать друг друга, проявляться в виде настойчивой тенденции, а случайные будут "гасить" друг друга, так как у разных элементов они разнонаправлены.

Одним из важных средств квантитативно-системного моделирования вероятностных систем, как это подчеркивает автор монографии, являются распределения. Самое общее определение распределения, даваемое автором диссертации, включает в себя

понимание этого объекта как упорядоченной совокупности количественно выраженных признаков объекта. "Важно не только то, что распределение выражает наличие внутренней упорядоченности в системе, но и то, что оно отражает взаимодействие между элементами и общность в их поведении, т.е. целостность системы, а также устойчивость и регулярность в массе вероятностно-случайных событий"(с. 40).

В работе разрабатывается общая типология видов распределений. Это один из ее важных методологических и методических моментов. Момент устойчивости, регулярности в вероятностных системах проявляется, с точки зрения автора монографии, и в устойчивости, относительной стабильности частот отдельных элементов или групп элементов.

Устойчивые, повторяющиеся моменты в организации различных подсистем лексики в анализируемой работе рассматриваются и с помощью нескольких методов группировки (классификации, кластеризации).

Необходимо особо отметить, что онтологической, качественной основой для определения и членения предмета исследования в данной работе, для выделения наиболее подходящих квантитативных методических средств и их аспектов является общая модель речевой деятельности. Для этого, в I-ой главе анализируется понятие речевого процесса со стороны таких компонентов как исходный замысел, языковой механизм, речевой продукт. Особое внимание уделяется таким разновидностям речевого продукта, как текст, и такому компоненту языкового механизма, как словарь.

Рассматриваемые в разделе I-ой главы единицы и уровни анализа (слово, словоупотребление, лексема, лексико-семантический вариант, классы лексем: лексико-семантические группы, лексико-грамматические, лексико-стилистические, полисемические, этимологические и т.д.) также онтологически обосновывают производимый далее квантитативный анализ.

На основе онтологического анализа объекта исследования и наиболее фундаментальной его аспектации ("язык - речь", "статика - динамика") выделяются и основные типы распределений языковых единиц (теоретические и эмпирические, динамические и статические).

Приложение таким образом обоснованных принципов моделирования к историческим и синхроническим проблемам организации лексики разных языков (эстонского, английского, русского, венгерского и др.) показало их высокую эффективность и

перспективность для последующего более широкого применения.

В результате анализа генетического состава эстонской лексики была выявлена подчиненность закону Вейбулла распределения древних элементов в словаре современного языка. Была также уточнена модель М.В. Арапова и М.М. Херц[+] связи между частотой употребления и возрастом слов.

В связи с анализом исторической динамики словаря было выявлено, что его рост происходит по логистическому закону, т.е. по одному из основных законов развития самоорганизующихся систем. При этом рассматривается вопрос о взаимной связи экспоненциального, логарифмического и логистического законов развития.

При анализе функционально-стилистической отмеченности лексики современного эстонского языка было выявлено, что численности различных функциональных стилистических помет распределяются в словаре по логарифмическому закону. Это - один из самых оригинальных и красивых результатов данного исследования.

Из качественных результатов анализа этого материала необходимо упомянуть, что Ю.А. Тулдавой выявлено, что разговорные и нейтральные лексические элементы в эстонском языке менее противопоставлены, чем, например, в русском. Однако причина такого положения еще не выявлена. В чем здесь дело: в разных путях развития литературных форм этих языков или во влиянии на это соотношение типа языков? Эти вопросы еще ждут своего решения.

Нуждается в своем объяснении и вопрос о значительно более высоком удельном весе в эстонском большом словаре стилистически отмеченной лексики в сравнении с большим русским словарем. На наш взгляд, причина в слабом отражении в большом русском словаре терминологической лексики[++], а она и является одной из главных разновидностей стилистически отмеченной лексики.

В связи с тем, что этот дефицит терминологической лексики в большом (семнадцатитомном) русском словаре, по нашим данным, приводит к определенному искажению и общего полисемического распределения в нем, к дефициту однозначных слов

[+] См.: Арапов М.В., Херц М.М. Математические методы в исторической лингвистике. М.: Наука, 1974. - 168 с.

[++] См.: А.А. Поликарпов. Полисемия: системно-квантитативные аспекты // Квантитативная лингвистика и автоматический анализ текстов (Уч. зап. Тартуского гос. ун-та. Вып. 774). Тарту, 1987. - С. 148-149.

в его структуре, можно сделать вывод, что терминологическая лексика входит составной необходимой частью в совокупный литературный словарь.

Особое значение имеют представленные в рецензируемой книге соотношения объемов групп слов разной полисемии. Параметры полученного распределения позволяют сравнивать между собой данные разных языков (венгерский, английский, русский) и делать типологически значимые выводы.

К числу важных проблем в рецензируемой книге относится исследование зависимости полисемических характеристик слов, их длины, словообразовательной активности и т.п. от частотных характеристик слов. Здесь автором выявляются важные изоморфизмы в законах связи этих характеристик, выявляется их комплексная взаимозависимость. Эти разделы исследования затрагивают важные в идейном и прикладном отношении вопросы устройства языка как оптимальной системы, системы с внутренней согласованностью ее важнейших характеристик+.

Проблемы частотных закономерностей употребления лексики в текстах представляют, несомненно, сердцевину анализируемой работы. Много внимания уделяется как выявлению частотных характеристик конкретных текстов (с целью их сравнения, определения степени их сходства, их лексико-стилистических характеристик и т.п.), так и выявлению общих закономерностей, управляющих частотными структурами лексики. В частности, выявляется лингвистико-типологический и стилистический смысл некоторых параметров аналитического описания частотной структуры.

Особое внимание автора диссертации привлек так называемый закон Ципфа. Рассматриваются различные его варианты, в т.ч. и самые новейшие, ищется между ними взаимосвязь и общая содержательная основа. По мнению Ю.А. Тулдавы, частотный закон Ципфа может быть выражением оптимальности (целесообразности, экономичности) использования языковых средств. Эта оптимальность, по его мнению, является результатом естественного процесса эволюционного развития самоуправляемых систем, связана с некоторыми фундаментальными функциями мозга, а также, видимо, с еще более общими, универсальными

+ Этот раздел книги Ю.А. Тулдавы перекликается по ряду своих идей с исследованиями Г. Альтмана, Р. Кёлера и некоторых других исследователей Бохумского университета (ФРГ). В этой связи см., например: R. Köhler. Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. - Bochum: Brockmeyer, 1986.

для мира живого (мира самоуправляемых систем) законами.

Вместе с тем, разумеется, в понимании даже общего смысла закона Ципфа, необходимых условий его реализации на лексическом материале продолжает оставаться много неясного. Обобщающие исследования типа того, которое предпринято в данной диссертации, продвигают нас по пути более глубокого проникновения в его суть.

Одной из важных сторон динамики частотной структуры является зависимость роста объема словаря от роста объема текста. В анализируемой работе эта зависимость исследуется на фоне некоторых глубинных факторов порождения речи, в частности, на фоне обнаруженного Ю.А. Тулдавой "давления рекуррентности", усиливающегося с ростом объема текстовой выработки. Предлагается и обсуждается математическая модель связи между объемом словаря и текста, основанная на принципе "ограничения разнообразия лексики".

Работу характеризует богатство идей и выводов. Вместе с тем для ее автора характерно стремление к поиску общих закономерностей, инвариантов, изоморфизма структур законов, управляющих организацией и функционированием лексики в тех или иных аспектах. В частности, в работе показана сводимость ряда специальных зависимостей к обобщенному статистическому закону Вейбулла.

В работе реализуется новый подход - квантитативно-системный - к анализу языковых фактов. Он позволяет значительно расширить собственно лингвистическую проблематику исследования систем языка и установить такие закономерности организации и функционирования языка, которые другим способом обнаружить невозможно. Количественный анализ в сочетании с качественным позволяет более глубоко познать лингвистический объект во всей его многогранности.

А.А. Поликарпов

Review of PROBLEMS AND METHODS OF THE QUANTITATIVE-SYSTEMIC INVESTIGATION OF VOCABULARY by Juhan Tuldava

Anatoli Polikarpov

S u m m a r y

The monograph under review presents a new synthetic approach to the study of language from the quantitative point of view. The fundamental theoretical and methodological principles and concepts of the quantitative-systemic analysis of vocabulary are discussed. Classification and modelling of the material in the form of statistical distributions are regarded as the principal methods of description and interpretation of linguistic data. Due attention is paid to the combination of quantitative approach with qualitative analysis. In four chapters the main laws of the statistical organization of text and vocabulary are discussed, the phonetical, grammatical, and semantic aspects of the investigation of vocabulary from the quantitative-systemic point of view are considered, some quantitative laws of diachronic lexicology are examined, and finally, the stylistic aspect of the functioning of vocabulary is discussed. The illustrative material has been taken from Estonian, Russian, English, German and other languages.

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА

Рецензия на монографию М.В. Арапова "Квантитативная лингвистика". - Москва: Наука, 1988. - 184 с.

Рецензируемая монография обобщает многолетние плодотворные исследования М.В. Арапова в области квантитативной лингвистики. Как отмечает автор во введении, основной целью, которую он преследовал при написании книги, являлся поиск связи между количественным и качественным аспектами в описании языка. Книга рассчитана на две профессиональные группы читателей: лингвистов, получивших традиционную подготовку, и на специалистов, в основном с базовым физико-математическим или техническим образованием, решающих комплекс задач, связанных с "человеко-машинным интерфейсом". Автор делает попытку сблизить методологическое понимание проблем разрабатываемых этими группами специалистов. Для первой из них книга полезна с точки зрения приобретения навыков приписывания качественным суждениям о языке точного количественного смысла. Второй группе читателей она может позволить более тесно связать решаемый ими круг задач с традиционными проблемами лингвистики. Предполагается, что читатель не имеет специальной математической подготовки и знаком с основами математического анализа, теорией вероятностей и математической статистики лишь в рамках обычного курса высшей математики, читаемого в технических вузах.

Монография состоит из введения, восьми глав, в которых лексика языка рассматривается с различных редко совмещающихся в языковедческих работах точек зрения, и заключения.

После краткого введения в первой главе автор подробно останавливается на значении количественных данных для изучения языка. Им убедительно показано, что любой лингвист, "если он выступает как практик, обязательно рассматривает язык не только в качественном, но и количественном аспекте". Далее подробно анализируются трудности, возникающие при попытках описать количественные закономерности языка в рамках системы представлений, сложившихся еще в шестидесятые годы и известных под названием лингвостатистики. В качестве альтернативного подхода М.В. Арапов развивает концепцию ценности языковых явлений, причем ценность им трактуется не просто в прагматическом смысле, а как неотъемлемый атрибут системы языка, как мостик, соединяющий качественную и коли-

чественную стороны его описания. Исходя из постулата, что на некоторых языковых объектах определено "не сводимое ни к какому другому структурное отношение ценностной упорядоченности", автор в дальнейшем подробно останавливается на анализе лексического уровня, для которого статус (ценность) того или иного элемента словаря определяется его местом в упорядоченном по убыванию списке слов, полученных путем обработки достаточно большой выборки текстов.

Вторая глава монографии посвящена рассмотрению частоты как характеристики употребительности слова в фиксированном тексте. Исходя из вариационного принципа, согласно которому реально наблюдаются лишь те из частотных структур, которым соответствует максимальное число различных способов их реализации, автор приходит к закону Ципфа и подробно останавливается на анализе дискретного варианта этого распределения. Необходимо особо подчеркнуть, что при своем обосновании закона Ципфа М.В. Арапов впервые обратил внимание на существование дополнительного отношения эквивалентности, возникающего при классификации элементов любого множества. Именно учет наличия этого дополнительного отношения, сформулированный автором как принцип диссимметрии системы, позволил ему получить выражение для функционала, экстремум которого и приводит к распределению Ципфа. Повидимому возможности этой плодотворной идеи еще далеко не исчерпаны и ее дальнейшее развитие позволит объяснить те реальные отклонения от идеальной модели, которые обсуждаются в конце главы при рассмотрении весьма обширного эмпирического материала.

В третьей главе, посвященной изменению употребительности слов в синхронии, развивается модель, учитывающая изменчивость статуса слов при переходе от одного текста к другому. Здесь подробно описывается методика сопоставления различных частотных словарей, позволяющая выразить сходство между этими словарями в количественной форме, рассматриваются различные числовые характеристики системы ранговых распределений двух сравниваемых словарей. В качестве меры расстояния между ними предлагается использовать разность логарифмов рангов слов в сопоставляемых словарях. Описанная методика может применяться как для выделения конкретной лексики, значимо различной по своим частотам употребления в сравниваемых текстах, так и для определения степени сходства лексического состава этих текстов в совокупности.

Связь употребительности слов с такими их характеристиками как возраст, длина, полисемия рассматривается, соответственно, в четвертой, пятой и шестой главах монографии. Последовательно развивая концепцию ценности (статуса) отдельных элементов словаря, М.В. Арапов в рамках достаточно естественных вероятностно-статистических моделей анализирует, как эти характеристики "в среднем" зависят от ранга слов, который используется в качестве меры их употребительности. Здесь приводятся оценки скорости изменения словаря языка в диахронии, найдены коэффициенты уравнений регрессии, связывающих среднюю длину и полисемию слов с их приведенным статусом (логарифмом ранга). Значительное место уделено рассмотрению количественной модели слова как сложного образования, состоящего из отдельных слогов, для описания структуры которых в свою очередь предложена интересная математическая модель. Отдельно анализируется дисперсия рассматриваемых параметров, что позволяет не только проверить статистическую значимость публикуемых результатов, но и служит дополнительным доказательством справедливости исходных положений развиваемой теории.

В заключительных главах монографии автор, опираясь на понятие ценности отдельного слова, переходит к обсуждению ценности определенных классов слов. Им предлагается способ определения продуктивности того или иного класса слов, из которого для непродуктивных классов вытекают такие свойства как уникальность, архаичность и высокая ценность в системе языка, тогда как, наоборот, продуктивные классы характеризуются массовостью, новизной и относительно низкой ценностью их элементов. В качестве меры продуктивности некоторого класса рассматривается относительное количество слов этого класса в упорядоченном словаре фиксированного объема. При этом, исходя из гипотезы о потенциальной бесконечности словаря класс определяется как непродуктивный, если число составляющих его слов принципиально конечно. В противном случае класс по определению считается непродуктивным.

Последняя глава книги посвящена изучению классов n-ок (кортежей) слов. В ней исследуется связь между количественными характеристиками форм, связанных парадигматическими отношениями. Основное внимание уделяется однородным бинарным категориям и приводится методика проверки категорий на однородность. Для однородных категорий устанавливается мера, позволяющая, зная частоту одного из членов пары, построить доверительный интервал для частоты встречаемости второго эле-

мента. При этом о регулярности той или иной категории можно судить, анализируя частоты всей совокупности пар, образующих данную категорию, исходя из числа элементов, попадающих в указанные интервалы.

Весьма интересно, что в качестве одного из основных выводов, полученных в результате проведенных исследований, М.В. Арапов приходит к своеобразному принципу неопределенности, согласно которому чем точнее мы знаем одни характеристики слова, тем другие характеристики этого же слова оказываются менее определенными. Насколько далеко простирается эта аналогия с известным в физике принципом неопределенности Гейзенберга и получит ли указанный принцип в будущем количественное выражение, могут показать лишь дальнейшие исследования.

Монографию в целом характеризует высокий уровень строгости изложения, четкое определение используемых понятий и математических конструкций. Каждая глава начинается с детальной постановки задачи и заканчивается подробным обсуждением интерпретации полученных результатов. Все изложение ведется в широком контексте сопоставления теоретических моделей с обширным эмпирическим материалом. Автором проделана большая работа по обработке необходимых справочных источников, много внимания уделяется чисто лингвистическим аспектам проблемы, принципиальным трудностям, возникающим при развитии теории и ее сопоставлении с весьма разнородными не всегда имеющимися в достаточном объеме экспериментальными данными. Значительное место в работе отводится обсуждению концепций и результатов опубликованных ранее другими исследователями.

Не все положения, высказанные автором, абсолютно бесспорны. В частности является дискуссионным высказанный на стр. 31 тезис о том, что в "целостном тексте лексическое разнообразие слов существенно больше, чем в конгломерате текстов того же объема". Могут вызвать возражения и отдельные упрощения (например, независимость структуры последующего слога в слове от предыдущего), положенные в основу тех или иных теоретических построений. Однако, как подчеркивает сам автор, он не столько стремился построить законченную по возможности более точную теорию, сколько ставил своей целью в рамках единого подхода установить "прозрачные логические связи между измеримыми характеристиками языка". Конечно, поставленная цель не могла быть достигнута без неминуемых

упрощений, которые отнюдь не умаляют достоинств рецензируемой монографии.

Книга в целом является безусловной удачей М.В. Арапова, отражает его высокую эрудицию, богата идеями и несомненно внесет весомый вклад в дальнейшее развитие квантитативной лингвистики.

Ю.К. Крылов

Review of QUANTITATIVE LINGUISTICS by M.V. Arapov

Yu.K. Krylov

S u m m a r y

The monography under review gives a full account of the fundamental principles of the quantitative approach to the study of language. Numerous ways are suggested of establishing links between findings of qualitative linguistic studies, on the one hand, and results obtained by methods of mathematical statistics, on the other. The key concept - the information value of linguistic units - is gradually and consistently evolved throughout the work. In particular, the relations of this value with such overt and measurable characteristics of words as their frequency, degree of polysemy, length, etc., are carefully traced.

The book is, without question, a major success. Specialists in a broad range of research areas, from general and applied linguistics to artificial intelligence, automatic text processing and information retrieval, will find it both useful and stimulating.

ОСНОВЫ СТИЛЕМЕТРИИ

Рецензия на книгу: Г.Я. Мартыненко. Основы стилеметрии. - Ленинград: Изд-во Ленингр. ун-та, 1988. - 174 с.

В рецензируемой книге дано общетеоретическое обоснование и убедительно продемонстрированы практические возможности новой филологической дисциплины - стилеметрии, которую автор рассматривает как часть стилистики и "которая занимается счетом и измерением стилистических явлений с целью упорядочивания и систематизации текстов и их частей" (С. 3). Стилеметрическая обработка текстов позволяет решить целый ряд практических задач, связанных с датировкой, периодизацией, диагностикой и атрибуцией текстов, а также дает возможность построить типологию текстов и их частей.

Стилеметрическая концепция автора выгодно отличается от работ прошлых лет, ориентированных в основном на определенные виды текстов. В рецензируемой книге разработаны универсальные принципы стилеметрического исследования, пригодные для анализа любых текстов - от поэтических до официально-деловых, - и для решения любой из вышеназванных практических задач. Важным достоинством предлагаемой концепции является также серьезный математический аппарат.

Основополагающие теоретические понятия стилеметрии рассмотрены в главе I. Автор обосновывает здесь филологический статус стилеметрии в связи с той ролью, которую играет в филологии понятие стиля. Детально рассмотрены существующие интерпретации этого понятия, в том числе и предлагаемая автором стилеметрическая интерпретация, в соответствии с которой стиль понимается как "совокупность измеримых симптоматических признаков, факультативно присутствующих в архетипе текста (текстов) и образующих периферию его характеристики, взаимосвязанную с глубинными, латентными характеристиками" (С. 55). В связи с этим показано, что̀ надо понимать под измерением стиля и в чем оно состоит. Проведены также аналогии между стилеметрией и такими "метрическими" науками, как биометрия и наукометрия. Для всех этих наук основным является статистический метод. Именно он определяет характер стилеметрического знания, процесс получения которого автор иллюстрирует наглядной схемой (см. С. 15).

Объектом стилеметрии, по словам автора, является текст, а ее предметом - наборы его существенных признаков, которые характеризуются мерой и количеством и тем самым допускают применение статистического метода. Поскольку со статистической точки зрения текст - это реальная совокупность, то в книге подчеркивается исключительная роль распределений и приводится их оригинальная типология, построенная по принципу 14 бинарных оппозиций типа ранг - частота, гауссовость - негауссовость, однородность - неоднородность и др. Детально рассмотрены также различные виды упорядочивающе-систематизирующих работ, которые приходится выполнять исследователю-стилеметристу.

Главы 2 и 3 имеют более специальный характер и требуют от читателя хорошей подготовки соответственно в области математической статистики и литературоведения.

В частности, в главе 2 обсуждаются проблемы так называемых статусных распределений, которые широко используются в экономике, социологии, демографии, науковедении, информатике, лингвистике, культурологии, географии, биологии и др. науках, когда изучаемые в них объекты упорядочиваются по их функциональному весу. В связи с этим возникает проблема моделирования таких распределений. В качестве материала, на котором автор иллюстрирует свои положения, используется частотный словарь предикатных слов из формул отечественных изобретений. Выбор этой лексики автор обосновывает жесткостью текста формулы изобретения, что находит свое выражение в ограниченности состава предикатных слов и в их высокой концентрации в верхней зоне частотного списка. Это создает благоприятные условия для анализа реального характера статусных распределений. Неоднородность этих распределений в данном случае, равно как и в других случаях, например, в случае распределений эпитетов с префиксом "без-" в поэзии Ф. Сологуба, В. Брюсова, К. Бальмонта и А. Блока, - нашла свое выражение в высокой степени вариации, сложном геометрическом характере кривой распределения, а также в неравномерном росте скользящего коэффициента вариации.

Автор книги приходит к чрезвычайно важному выводу о том, что наиболее подходящими видами статистик статусных распределений являются ранговые статистики, поскольку они слабо зависят от объема выборки и обеспечивают возможность надежного измерения концентрации лексико-семантических данных в тексте и корпусе текстов. С этой целью в работе приво-

дится детальное и корректное описание математических свойств ранговых статистик, что можно рассматривать как важный вклад в математическую статистику.

Хотя основные результаты данной главы получены на основе лингвистических данных лексико-семантического характера, они имеют междисциплинарную значимость, так как относятся ко всем приводившимся выше наукам. Это обусловлено, по словам автора, тем, что "закономерности концентрации и рассеяния единиц в организованных совокупностях самой разнообразной природы имеют сходный характер" (С. 105).

В главе 3, посвященной стилеметрическому анализу литературно-художественных систем, особый интерес представляет вопрос об индивидуальном и коллективном в литературном творчестве. Это естественно, поскольку именно в беллетристике с особенной яркостью проявляется авторская индивидуальность, находящая свое выражение прежде всего в индивидуальном стиле, в связи с чем возникает вопрос, можно ли в такой ситуации говорить о некоторых общих чертах, присущих текстам множества литераторов. Автор рецензируемой монографии отвечает на этот вопрос положительно. Коллективное в художественных текстах, по его мнению, проявляется в том, что любое произведение "создается в единой культурно-исторической среде, на фоне языковых привычек, вырабатываемых стихией зримых и незримых контактов всех писателей, живущих в данную или хронологически смежных эпохах" (С. 107). При этом применение принципов массового анализа к изящной словесности нисколько не принижает ее, поскольку "яркая стилистическая индивидуальность, - как справедливо отмечает автор, - только выиграет, если мы будем рассматривать ее на фоне обыденной массы" (С. 108).

При определении наиболее релевантных признаков художественного стиля автор, как представляется, вполне обоснованно отдает предпочтение синтаксису, соглашаясь с мнением тех исследователей, которые утверждают, что "ключ к слогу писателя - в его синтаксисе". В подтверждение этого тезиса приводится ряд убедительных аргументов.

Особое внимание в этой связи автор уделяет измерению синтаксической сложности. В книге разграничивается иерархическая и линейная сложность. Та и другая оцениваются такими параметрами, как ширина дерева и его длина, мощность уровня и интегральная ширина, степень гнездования, степень разрывности, степень дистантизации и др.

Предложенные автором синтаксические параметры (6 основных и 3 дополнительных) были использованы для анализа текстов 100 писателей XIX и XX вв. В результате была получена многомерная классификация этих авторов, т.е. разбиение на их однородные группы по каждому из предложенных параметров.

Резюмируя, можно с полным основанием сказать, что в монографии Г.Я. Мартыненко дана всесторонняя разработка стилеметрии как целостной науки с собственным исследовательским аппаратом, базу которого образуют статистические методы, и продемонстрированы практические возможности этой науки, которые открывают широкие перспективы для решения самых различных филологических задач.

Вместе с тем следует отметить, что при характеризации стиля все признаки, которыми пользуется автор, относятся к поверхностному синтаксису. Это, естественно, значительно упрощает операции счета и измерения, но не исключено, что наряду с этим упрощает и суть дела. Автор, правда, говорит о глубинных, латентных характеристиках стиля, отражаемых его поверхностными параметрами. Однако прямой связи между ними не наблюдается, тем более если учесть тот факт, что о синтаксических тонкостях стиля автор трактует в терминах достаточно примитивной грамматики зависимостей. В принципе можно было бы, минуя поверхностный уровень, непосредственно обратиться к глубинно-синтаксической сфере. Это можно было бы сделать на базе трансформационного синтаксиса, в терминах которого каждое поверхностное предложение может быть охарактеризовано числом входящих в него ядерных предложений и примененных к ним трансформаций.

Стиль художественных произведений характеризуется также некоторыми текстуальными параметрами типа нарративности, дескриптивности или аргументативности, указывая на преобладание в тексте сменяющих друг друга эпизодов, описания природы или обстановки или же рассуждений на ту или иную тему. С этим связано также и то, что преимущественно описывается в тексте – события или факты. Можно было бы выделить и другие стилистически существенные факторы или другие синтаксические концепции, в терминах которых можно было бы более всесторонне характеризовать стиль. Такой анализ мог бы приблизить исследователя к тем латентным стилистическим характеристикам, о которых пишет автор. Но вместе с тем, он в громадной степени усложнил бы выделение считаемых и измеряемых единиц.

Сказанное здесь не следует рассматривать как претензию к книге Г.Я. Мартыненко. Это всего лишь возможные наметки для дальнейших поисков.

Что же касается самой монографии, то в ней обобщен и синтезирован большой лингвистический, литературоведческий и статистический материал, на базе которого автору удалось получить качественно новое знание. Несомненно, что специалисты с большим интересом встретят содержательное и стимулирующее исследование Г.Я. Мартыненко.

В.В. Богданов

Review of FUNDAMENTALS OF STYLOMETRY by G.Y. Martynenko

Valentin Bogdanov

S u m m a r y

The monograph under review deals with some basic concepts and applications of stylometry, a new science which is considered to be of great value for a variety of philological investigations, especially for evaluation of literary styles.

The monograph consists of 3 chapters. Chapter 1 includes a system of stylometric principles and a detailed description of statistical distributions on the basis of 14 binary oppositions such as rank - frequency, construction - behavior, element - set of elements, uniformity - nonuniformity, stability - nonstability, symmetry - asymmetry, statics - dynamics, etc.

Chapter 2 deals with the so-called status distributions widely used in many sciences such as economics, sociology, linguistics, information science, geography, biology etc., if their elements are arranged according to their functional weight, e.g. lexical items may be arranged in accordance with their occurrence in a text.

Chapter 3 presents a comparative description of the texts of 100 Russian writers of the 19-th and 20-th centuries in terms of 9 syntactic features such as a number of nested constructions, the length and width of sentence tree-like diagrams, sentence size, etc. As a result the author managed to obtain a multi-dimensional classification of the 100 Russian writers on the basis of the above-mentioned syntactic features.

ВОЛНОВАЯ ТЕОРИЯ КАК ОДНО ИЗ НАПРАВЛЕНИЙ КВАНТИТАТИВНОЙ ЛИНГВИСТИКИ СВЯЗНОГО ТЕКСТА

(Обзор-рецензия работ Ю.К. Крылова)

А.А. Поликарпов, Ю.А. Тулдава

Среди большого числа работ, посвященных изучению статистических закономерностей организации лексики в целостных текстах, особое место занимает разрабатываемая Ю. К. К р ы л о в ы м волновая теория (Yu.K. Krylov, M.D. Yakubovskaya, 1983; 1982; 1985; Ю.К. Крылов 1982 - 1988).

Первоначально идея перенесения квантово-механического дуализма волна-частица на дихотомию язык - речетворческий процесс (Yu.K. Krylov, M.D. Yakubovskaya, 1983) воспринималась большинством исследователей скорее как экстравагантная аналогия, чем перспективный подход к проблемам порождения связного текста. Скептическое отношение к этому подходу объяснялось господствующей точкой зрения, что такой сложный и противоречивый объект как "речевая деятельность" существенно отличается в онтологическом плане от систем, изучаемых естественными науками. Однако, по мере того как теория развивалась (Крылов Ю.К., 1986, 1987, 1988), становилось все более очевидным, что разрабатываемая концепция вполне вписывается в общий квантитативно-системный подход к изучению лексики. В настоящее время можно считать уже общепризнанным, что актуальность исследований в области волновой теории порождения текста связана не только с проблемами теоретического и прикладного языкознания, но и в не меньшей степени обусловлена все возрастающими запросами практики к общенаучному изучению любых сложных самоорганизующихся систем. В этой связи успехи, достигнутые в области применения формализма квантовой механики к описанию статистических связей между единицами различных уровней организации речетворческого процесса, представляют определенный интерес и с точки зрения общефилософских проблем, логики и методологии разработки обобщенной квантовой теории.

Следует отметить, что идея возможности использования результатов современной квантовой физики при построении квантитативной теории языка последовательно проводится Ю.К. Крыловым начиная с его первых публикаций. Так еще в работе (Крылов Ю.К., 1982) он писал: "Трудно представить, что статистический механизм, ответственный за возникновение распределений типа распределения Ципфа - Мандельброта, находится

вообще вне парадигмы статистик теоретической физики". Замена классического полиномиального распределения на распределения, подчиняющиеся статистикам Ферми-Дирака и Бозе - Эйнштейна позволила ему объяснить S-образный характер кривой, описывающей частоты встречаемости букв и фонем в естественных языках (Крылов Ю.К., 1982), предложить теоретические формулы, достаточно хорошо аппроксимирующие полисемические распределения. (Yu.K. Krylov, M.D. Yakubovskaya, 1982). Наконец предтечей создания собственно волновой теории следует считать, по-видимому, сразу две работы (Yu.K. Krylov, M.D. Yakubovskaya, 1983; Лебедев А.Н., 1983), появившиеся практически одновременно в конце 1983 года, в которых авторы обратили внимание на то, что классический закон Ципфа $F = C i^{-1}$ получается почти автоматически, если предположить, что с порождением текста связан периодический процесс с набором частот, соотносящихся как члены гармонического ряда: 1, 1/2, 1/3, 1/4 ....

Общие положения развиваемой теории сформулированы Ю.К. Крыловым в работах (Крылов Ю.К., 1987; 1988). Отдельные модели неоднократно обсуждались на семинарах группы "Текст как объект междисциплинарных исследований" (28-30.01. 1986, Тарту; 23-29.04. 1987, Боржоми - Тбилиси; 3-8.10. 1987, Звенигород; 1-5.11. 1988, Тбилиси), докладывались автором на всесоюзном школе-семинаре "Психологическая бионика" (22-24.04. 1986, Харьков; 18-23.05. 1987, Харьков; 11-15.05. 1988, Харьков) и на У симпозиуме по лингвистическим проблемам искусственного интеллекта 11-13.04. 1988, Ленинград).

В окончательном варианте теории порождение текста рассматривается как многомерный случайный процесс, развивающийся в пространстве счетного числа допустимых состояний. Каждое состояние характеризуется своей собственной функцией $\psi_k(x)$, где x - непрерывная координата (адрес) очередного словоупотребления $/x \in [0, N]$; N - полная длина текста. Постулируется, что все вхождения в текст слов с фиксированной кратностью k относятся к одному и тому же k-му состоянию. При этом вероятность того, что k-кратное слово будет принадлежать именно отрезку текста от x до $x + dx$ дается элементом вероятности $\psi_k^2(x)\,dx$.

Теория основывается на известной теореме о том, что любая квадратично интегрируемая функция $\psi(x)$ допускает разложение в обобщенный ряд Фурье по произвольной ортонормированной системе собственных функций $\{\psi_k(x)\}$. Для системы,

обладающей свойством полноты, коэффициенты линейной комбинации

$$\psi(x) = \sum_k C_k \varphi_k(x) \qquad (1)$$

удовлетворяют уравнению замкнутости:

$$\int_0^N \psi^2(x)\,dx = \sum_k C_k^2 . \qquad (2)$$

Соотношение (2) позволяет интерпретировать C_k^2 как вероятности "возбуждения" соответствующих состояний, а $C_k^2 \varphi_k^2(x)dx$ как безусловные вероятности принадлежности k-кратных слов интервалу $[x, x+dx]$. Соответственно, безусловная вероятность появления в этом интервале любого слова словаря текста (локальная плотность покрытия текста словарем) пропорциональна $\psi^2(x)dx$.

Утверждается, что все статистические свойства текста уже фиксированы, если задана его "вольновая" функция $\psi(x)$, зависящая лишь от адресов вхождения в текст элементов словаря. В условиях стационарности процесса порождения текста $\psi(x) = Const$, что с учетом нормировки $\int_0^N \psi^2(x)dx = 1$ приводит к соотношению $\psi(x) = \frac{1}{\sqrt{N}}$.

Принимается, что математические ожидания m_k числа k-кратных слов в тексте пропорциональны объему его словаря $\mathcal{L}$ с коэффициентами пропорциональности C_k^2, т.е. $m_k = C_k^2 \mathcal{L}$.

Следует подчеркнуть, что одним из достоинств рецензируемого направления является открытость теории к построению различных моделей текстообразования, отличающихся конкретным выбором системы собственных функций $\{\varphi_k(x)\}$. После того как в рамках той или иной модели последние уже определены, для вычисления C_k достаточно умножить (I) на $\varphi_k(x)$ и проинтегрировать в интервале от 0 до N. В результате

$$C_k = \int_0^N \psi(x)\,\varphi_k(x)\,dx . \qquad (3)$$

Для описания лексического (частотного) спектра в качестве $\{\varphi_k(x)\}$ Ю.К. Крылов предлагает использовать собственные функции оператора Гамильтона:

$$\hat{H} = -\frac{\hbar^2}{2} \sum_\ell \frac{\Delta_\ell}{m_\ell} + U(x_1 \ldots x_\ell \ldots), \qquad (4)$$

где $U(x_1 \ldots x_\ell \ldots)$ - потенциальная функция, зависящая как

от макроструктуры порождаемого текста, так и от характера взаимодействия составляющих его элементов, Δ_ℓ - известный оператор Лапласа, m_ℓ - индивидуальная характеристика ℓ-го элемента словаря, $\hbar$ - некоторая универсальная постоянная. При этом набор собственных функций оператора $\hat{H}$ должен удовлетворять уравнению Шредингера для стационарных состояний:

$$\hat{H}\psi = \lambda\psi . \qquad (5)$$

В качестве простейшего примера конкретизации общей теории Ю.К. Крыловым был рассмотрен вариант: $U(x, \ldots x_\ell) \equiv 0$. В этом случае в качестве системы собственных функций выступает система тригонометрических функций $\psi_\ell(x) = A_\ell \sin\frac{\ell\pi x}{N}$. Из условий нормировки $\int_0^N \psi_\ell^2(x)dx = 1$ вытекает, что $A_\ell = \sqrt{\frac{2}{N}}$ в силу чего

$$C_\ell = \int_0^N \frac{1}{\sqrt{N}}\sqrt{\frac{2}{N}}\sin\frac{\ell\pi x}{N}\,dx = \begin{cases} 0 & \ell = 2k \\ C_k = \frac{2\sqrt{2}}{\pi(2k-1)} & \ell = 2k-1 \end{cases} \qquad (6)$$

Таким образом данная модель приводит к лексическому спектру вида

$$m_k = \frac{2\mathcal{L}}{\pi^2(k-0{,}5)^2} . \qquad (7)$$

Уместно отметить, что именно соотношения (7) использовались непосредственно Дж. Ципфом в качестве уточненного варианта для аппроксимации лексического спектра.

Еще более интересной представляется модель, предложенная Ю.К. Крыловым для описания лексических спектров коротких текстов ($N \leq$ 1000 словоупотреблений). В этой модели, известной в квантовой механике под названием уравнения Шредингера для потенциального ящика конечной высоты, собственные функции $\psi_k(x)$ определяются как множество допустимых решений уравнения:

$$\frac{d^2\psi}{dx^2} + \mu_0\left[E - U(x)\right] = 0, \qquad (8)$$

в котором $\mu_c = m\hbar$, $U(x) = \begin{cases} U_0 & x \leq 0 \quad x \geq N \\ 0 & 0 < x < N \end{cases}$

m и U_0 - параметры, отвечающие, соответственно: m - за выбор единиц подсчета словаря текста (лексемы, словофор-

мы, ЛСВ и т.п.), U_0 - за свойства конкретного текста.

Автором показано, что при условии равномерного покрытия текста словарем $(\psi(x) = Const)$ численности k-кратных слов должны удовлетворять соотношениям:

$$m_k = C_k^2 \mathcal{L} = 2\left(\frac{\eta_k}{\xi_k \mu}\right)^2 \frac{\eta_k}{\eta_k + 1} \mathcal{L}, \qquad (9)$$

Последняя модель фактически является однопараметрической, так как лексический спектр в ней зависит лишь от одного результирующего параметра

$$\mu = \frac{1}{2}\sqrt{\mu_0 U_0} N = \gamma N \qquad \left(\gamma = \frac{1}{2}\sqrt{\mu_0 U_0}\right). \qquad (10)$$

При этом условия ограниченности $\psi(x)$ на бесконечности и непрерывности их первых производных оставляют допустимыми лишь значения η_k и ξ_k, удовлетворяющие системе трансцендентных уравнений:

$$\eta_k^2 + \xi_k^2 = \mu^2; \quad \eta_k = \xi_k \operatorname{tg} \xi_k. \qquad (11)$$

Когда μ мало, система (II) имеет единственное решение при $k = 1$, что соответствует наличию в очень коротком тексте лишь одноразовых слов. По мере увеличения длины текста $\mu = \gamma N$ возрастает. Это приводит к увеличению числа существующих решений. Последовательно появляются решения при $k = 2, 3, 4 \ldots$ (в тексте начинают присутствовать двукратные, трехкратные и т.д. слова). В предельном случае $\mu \to \infty$ $(U_0 \to \infty)$ спектр асимптотически стремится к виду (7). Таким образом, указанная модель позволяет не только теоретически рассчитать спектр текста с тем или иным объемом словаря, но и предсказывает динамику изменения этого спектра.

Одной из интересных особенностей данной модели является также предсказываемое ею наличие так называемого "сплошного спектра" решений $\psi_\omega = A_\omega \sin(\omega x + \varphi_\omega)$, с малой вероятностью имеющих место при $\omega > \sqrt{\mu_0 U_0}$. Существование этих решений показывает, что всегда имеется отличная от нуля вероятность порождения текстов с "аномальными" спектрами (при желании автор может в некоторой мере нарушить системные отношения и создать текст с произвольно организованными повторами).

Другим, с нашей точки зрения интересным эффектом, яв-

ляется предсказание возможности появления текстов, спектр которых убывает не монотонно, а обладает локальными максимумами при определенных, достаточно больших k . Сопоставление теории с эмпирическим материалом показало, что указанный эффект действительно имеет место.

Следует также обратить внимание на то, что Ю.К. Крылову удалось впервые связать в рамках единой концепции связного текста такие наблюдаемые характеристики текста как его спектр и плотность покрытия словарем. В частности, используя при известном спектре (I) и полагая $C_k = \sqrt{m_k}$, теория позволяет рассчитать плотность покрытия. Наоборот, формулы (3) в рамках той или иной модели позволяют решать обратную задачу.

Оценивая предложенный Ю.К. Крыловым подход к вопросам порождения связного текста в целом, подчеркнем, что уже в настоящее время он дает значительные результаты. Еще более интересными представляются его перспективы. Следует пожелать автору дальнейших успехов на пути разработки предлагаемой им теории.

Л И Т Е Р А Т У Р А

Крылов Ю.К. Об одной парадигме лингвистических распределений // Учен. зап. ТГУ, вып. 628. - Тарту, 1982. - С. 80 - 102.

Крылов Ю.К. Статистическая организация текста и вариационные принципы // Материалы межвузовского семинара, 28.10 - 01.11. 1986. - Тбилиси, 1987. - С. 108 - 111.

Крылов Ю.К. Квантитативно волновой дуализм порождения связного текста // Прикладная лингвистика и автоматический анализ текста. - Тарту: Из-во ТГУ, 1988. - С. 46 - 47.

Крылов Ю.К. Ранговые распределения и элементы симметрии связного текста // Семинар "Текст как семиотический объект: качественные и количественные аспекты его организации". 28-30.01. - Тарту, 1986.

Крылов Ю.К. Стохастические модели порождения текста на базе квантово-механических представлений // Семинар "Квантитативные аспекты системной организации текста" 28.10 - 1.11. - Тбилиси, 1986.

Крылов Ю.К. Возможности оптимизации обработки текстовой информации при использовании предпочтений в распознавании графов различных структурных единиц текста // Все-

союзная школа-семинар "Психологическая бионика" 22.04 - 24.04. - Харьков, 1986.

Крылов Ю.К. Стационарная модель порождения текста и проблема его расчлененности // Семинар "Структурные единицы языка и текста: проблемы междисциплинарных исследований" 23.04-28.04. - Боржоми, 1987.

Крылов Ю.К. Нейрофизиологические предпосылки к построению стохастических моделей порождения и восприятия речи // Всесоюзная школа-семинар "Бионика интеллекта" 18.05 - 23.05. - Харьков, 1987.

Крылов Ю.К. Порождение текста как слабо стационарный случайный процесс // Семинар "Системно квантитативные проблемы исследования языка и текста" 3.10.-8.10. - Звенигород, 1987.

Крылов Ю.К. Анализ возможностей исследования семантической структуры связного текста формализованными методами // У симпозиум по лингвистическим проблемам искусственного интеллекта, 11.04-13.04. - Ленинград, 1988.

Крылов Ю.К. Порождение и восприятие речи как вероятностный процесс // Всесоюзная школа-семинар "Психологическая бионика", 11.05-15.05. - Харьков, 1988.

Крылов Ю.К. Начала волновой теории связного текста // Семинар "Системно-квантитативный анализ текста и языка: новые подходы и методы", 1.11-5.11. - Тбилиси, 1988.

Лебедев А.Н. Закономерности повторения слов в речи // Психологический журнал. Том 4, № 5, 1983. - С. 11 - 22.

Krylov Yu.K., Yakubovskaya M.D. On So-called Information Redundancy of Natural Semiotic Sequences // Symposium on Common Aspects of Processing of Linguistic and Musical Data: Summaries. - Tallinn, 1982, pp. 59 - 65.

Krylov Yu.K., Yakubovskaya M.D. On Some Possibilities of Applying the Quantum-mechanical Formalism when Constructing Stochastic Models of Text Generation and Recognition // Symposium on Grammars of Analysis and Synthesis and Their Representation in Computational Structures: Summaries. - Tallinn, 1983, pp. 49 - 51.

Krylov Yu.K., Yakubovskaya M.D. The Possibilities of a GLS Matrix in the Optimization of Algorithms for the Automatic Compilation of Frequency Dictionaries and for the Statistical Analysis of Coherent Texts // Symposium on Automatic Compilation of Dictionaries: Summaries. - Tallinn, 1985, pp. 29 - 30.

THE WAVE-THEORY - AN IMPORTANT TREND IN QUANTITATIVE LINGUISTIC STUDIES OF COHERENT TEXTS

Anatoli Polikarpov, Juhan Tuldava

S u m m a r y

The article presents a survey and analysis of Yu. K. K r y l o v ' s work in developing the so-called wave theory of texts, in which the classical "wave - particle" dualism of quantum mechanichs is applied to linguistics where it takes the form of a dichotomy "language - text generation process". The achievements of the new approach are discussed and a number of problems yet to be solved pointed out. The conclusion is reached that the approach has enabled Yu. K. Krylov to link up, in a novel theoretical framework, such manifest and measurable characteristics of a coherent text as its frequency spectrum, on the one hand, and the vocabulary size as dependent on text size, on the other.

С О Д Е Р Ж А Н И Е

ХРОНИКА:

РЕЦЕНЗИИ:

SUMMARIES

SURVEY:

REVIEW:

Ученые записки Тартуского университета.
Выпуск 872.
КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ
АНАЛИЗ ТЕКСТОВ 1989.
На русском языке.
Резюме на английском языке.
Тартуский университет.
ЭССР, 202400, г.Тарту, ул.Юликооли, 18.
Ответственный редактор Ю. Тулдава.
Подписано к печати 13.XI.1989.
Формат 60x90/16.
Бумага писчая.
Машинопись. Ротапринт.
Учетно-издательских листов 9,50. Печатных листов 10,75.
Тираж 550.
Заказ № 787.
Цена 1 руб. 90 коп.
Типография ТУ, ЭССР, 202400, г.Тарту, ул.Тийги, 78.
6 - 2