ISSN 0207-4435

TARTU RIIKLIKU ÜLIKOOLI TOIMETISED

УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS

711

KVANTITATIIVLINGVISTIKA JA TEKSTIDE AUTOMAATANALÜÜS

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

1985

TARTU 1985

TARTU RIIKLIKU ÜLIKOOLI TOIMETISED
УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS
ALUSTATUD 1893.a. VIHIK 711 ВЫПУСК ОСНОВАНЫ В 1893.г.

KVANTITATIIVLINGVISTIKA JA TEKSTIDE AUTOMAATANALÜÜS

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

1985

TARTU 1985

Kogumik "Kvantitatiivlingvistika ja tekstide automaatanalüüs" jätkab sarja "Töid keelestatistika alalt", mida ilmus 10 väljaannet (1976 - 1984). Käesolevas väljaandes (1985) on avaldatud Tartu Riikliku Ülikooli rakenduslingvistika rühma liikmete ja välisautorite artiklid.

Сборник "Квантитативная лингвистика и автоматический анализ текстов" является продолжением серии "Труды по лингвостатистике" (10 выпусков в период 1976 - 1984 гг.).Настоящий выпуск 1985 г. содержит статьи сотрудников Группы прикладной лингвистики при Тартуском государственном университете и исследователей из других городов.

The collections "Quantitative Linguistics and Automatic Text Analysis" continue the series "Papers on Linguo-Statistics" (published serially as issues of Acta et Commentationes Universitatis Tartuensis since 1976). The present issue (1985) contains investigations by members of the Research Group of Applied Linguistics at Tartu State University and by guest authors.

СТАБИЛЬНОСТЬ ЧАСТОТ НЕМЕЦКИХ ГРАФЕМ

Б.Н. Гвоздович

Графемы как единицы особого уровня языка, на котором "отношение между планом содержания и планом выражения наиболее детерминировано по сравнению с другими его уровнями" (Гусева Е.К., 1973, с. 202), в каждом языке образуют систему, элементы которой в силу особенностей данного языка встречаются в тексте с разной частотой. Поэтому проблема стабильности частот языковых единиц, которая занимает в настоящее время одно из центральных мест в лингвистической статистике, по отношению к графемам существует в двух аспектах. Во-первых, насколько стабильны частоты графем данного языка, если рассматривать сразу всю систему графем или какие-то ее структурные части. Во-вторых, насколько стабильны частоты каждой графемы в отдельности. Настоящее исследование посвящено этой последней задаче.

Исследование проводилось на выборке объемом в 50 000 графем (500 отрывков по 100 графем), составленной из произведений современных немецких писателей, а также из научных и газетных статей. Выборка была организована таким образом, что каждый из трех основных литературно-художественных жанров, а именно, проза, поэзия и драма, а также оба упомянутых выше функциональных стиля речи, тесно связанные с письменной формой языка, были представлены в ней разными долями по 100 отрывков (10 отдельных произведений)[+]. Внутри произведения отрывки располагались на все возрастающем удалении друг от друга (от 100 до 10 000 графем) и представляли разные его части - начало, середину и конец. Задачей исследования было выяснить, носят ли частоты немецких графем стабильный характер:

- внутри одного произведения;
- внутри всего массива исследованных текстов.

В результате подсчетов было получено 1194 частотных ряда, репрезентирующих статистическое поведение отдельно каждой немецкой графемы в указанных выше произведениях - по

[+] Словом "произведение" здесь и далее ради единообразия наряду с законченными произведениями (например, пьесами и научными статьями) называются также главы романов и большие части газетного текста (например, полосы).

50 рядов на каждую графему, кроме J, Q, X, Y, которые в некоторых произведениях не встретились ни разу, и поэтому их частоты образовали соответственно 49, 9, 14 и 22 ряда распределения. Задачи эксперимента решались на основе сопоставления этих эмпирических распределений с теоретическими.

Основным инструментом исследования служил критерий согласия "хи-квадрат", с помощью которого проверялась нулевая гипотеза, состоящая в предположении, что исследуемые эмпирические распределения хорошо согласуются с теоретическими. Теоретические частоты находились при этом по формуле

$$n_i^o = P_i N , \qquad (1)$$

где n_i^o - теоретическая частота i-ой графемы, P_i - ее же относительная частота, а N - объем текста. Значение χ^2 вычислялось по формуле

$$\chi^2 = \sum \frac{(n_i - n_i^o)^2}{n_i^o} , \qquad (2)$$

где n_i - эмпирическая частота i-ой графемы.

Принятие нулевой гипотезы при этих условиях означает, что расхождения между наблюдаемыми и теоретическими частотами данной графемы статистически незначимы, а, следовательно, эмпирические частоты можно рассматривать как закономерное варьирование некоей теоретической частоты (вероятности), постоянной для данной графемы. Поэтому в описываемом эксперименте принятие нулевой гипотезы считалось доказательством стабильности исследуемых частот, а ее отвержение - показателем нестабильности этих частот.

В соответствии с названными выше целями все исследование проводилось в два этапа. На первом этапе исследовался вопрос стабильности частот графем в рамках одного произведения. При этом относительные частоты графем P_i, необходимые для вычисления их теоретических частот, находились отдельно для каждого произведения по данным всех 10 его отрывков. В таблице 1 в качестве примера приведены эмпирические частоты (первая строка), теоретические частоты (вторая строка) и значения χ^2 (третья строка), полученные при исследовании стабильности частот графем в рамках одного произведения (Wolf Ch., 1963, с. 294 и след.).

Полученные значения χ^2 сравнивались с критическими значениями этого критерия при данном числе степеней свободы и 1%-ном уровне значимости: нелевая гипотеза принималась, если найденное значение χ^2 не превышало критического при ука-

занном уровне значимости, и отвергалась, если оно превышало критическое при этом уровне значимости (Урбах В.Ю., 1964, с. 359). Эти условия обеспечивают 99%-ную вероятность того, что отвергаемая нулевая гипотеза действительно ложна.

На этом этапе эксперимента нулевая гипотеза была отвергнута только в 2 случаях из 1194. Один из них относится к графеме F (в сборнике стихов "Новая немецкая лирика"), а второй - к графеме K (в романе Б. Келлермanна "Пляска смерти"). Во всех остальных случаях нулевая гипотеза была принята.

Эти результаты дают основание утверждать, что частоты всех немецких графем, кроме F и K внутри одного и того же произведения носят стабильный характер. Что же касается частот графем F и K, то в одних произведениях они стабильны, а в других (очень немногочисленных) - нестабильны. Интересно отметить, что все случаи нестабильности частот этих двух графем приходятся на произведения художественной литературы, в то время как в газетных и научных текстах они с этой точки зрения ведут себя так же, как и частоты других графем.

Таблица I

Стабильность частот графем A, C, L, N, Z
в рамках одного произведения

Графема		Отрывок текста										
		1	2	3	4	5	6	7	8	9	10	Σ
A	n_i	11	8	10	3	7	9	7	5	11	7	78
	n_i^0	7.80	7.80	7.80	7.80	7.80	7.80	7.80	7.80	7.80	7.80	78
	χ^2	1.31	0	0.62	2.95	0.08	0.18	0.08	1.00	1.31	0.08	7.61
C	n_i	5	4	4	2	1	1	-	3	2	1	23
	n_i^0	2.30	2.30	2.30	2.30	2.30	2.30	2.30	2.30	2.30	2.30	23
	χ^2	3.16	1.25	1.25	0.04	0.73	0.73	2.30	0.21	0.04	0.73	10.44
L	n_i	4	5	5	4	-	7	1	2	4	8	40
	n_i^0	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	40
	χ^2	0	0.25	0.25	0	4.00	2.25	2.25	1.00	0	4.00	14.00
N	n_i	10	6	8	6	10	8	11	9	9	11	88
	n_i^0	8.80	8.80	8.80	8.80	8.80	8.80	8.80	8.80	8.80	8.80	88
	χ^2	0.16	0.89	0.07	0.89	0.16	0.07	0.55	0	0	0.55	3.34
Z	n_i	1	-	2	2	-	1	1	3	2	1	13
	n_i^0	1.30	1.30	1.30	1.30	1.30	1.30	1.30	1.30	1.30	1.30	13
	χ^2	0.07	1.30	1.37	1.37	1.30	0.07	0.07	2.25	1.37	0.07	9.22

На втором этапе эксперимента изучался вопрос стабильности частот немецких графем внутри всего массива исследованных текстов. На этом этапе относительные частоты графем P_i, необходимые для вычисления их теоретических частот, находились по данным всего эксперимента, т.е. как доли текста длиной в 50 000 графем, и их поэтому можно рассматривать как относительные частоты (вероятности) немецких графем вообще. Эти относительные частоты приведены в таблице 2.

Таблица 2

Относительные частоты немецких графем

Графема	Относительная частота	Графема	Относительная частота	Графема	Относительная частота
A	0,05790	J	0,00292	S	0,07080
B	0,01936	K	0,01430	T	0,06036
C	0,03162	L	0,03686	U	0,04400
D	0,05118	M	0,02758	V	0,00820
E	0,16342	N	0,09640	W	0,01746
F	0,01644	O	0,02646	X	0,00054
G	0,02904	P	0,00772	Y	0,00070
H	0,04864	Q	0,00032	Z	0,01102
I	0,08398	R	0,07276	Всего	1.00000

Эмпирические ряды распределения, которые на этом этапе сравнивались с теоретическими, были составлены из частот, найденных для каждой графемы во всех 50 произведениях. Всего было исследовано 26 таких рядов по 500 величин в каждом. В таблице 3 приведены полученные при этом значения.

Таблица 3

Значения, полученные при проверке частот графем на стабильность на всем массиве текстов

Графема	Значение χ^2	Графема	Значение χ^2	Графема	Значение χ^2
A	418,58	J	548,97	S	452.72
B	450.96	K	508.31	T	413.56
C	555.37	L	479.38	U	424.86
D	385.12	M	561.07	V	484.65
E	296.21	N	357.34	W	536.92
F	555.88	O	505.17	X	531.30
G	515.56	P	536.10	Y	571.82
H	551.73	Q	659.71	Z	508.39
I	320.39	R	403.83		

При 499 степенях свободы и 1%-ном уровне значимости критическое значение χ^2 не больше 576,49[+]. Из приведенных в таблице 3 значений χ^2 только одно (у графемы Q превышает критическое. Следовательно, нулевая гипотеза отвергается только для ряда распределения частот этой графемы, в остальных же 25 случаях она не отвергается. Это означает стабильность частот всех немецких графем в рассмотренных условиях кроме графемы Q.

Нестабильность частот этой графемы объясняется следующим образом. Графема Q – самая редкая немецкая графема, она встречается в среднем 3 раза на 10 000 букв, поэтому в отрывке длиной в 100 графем ее обычно не бывает. В то же время в некоторых текстах в первую очередь в научных и газетных, в силу их тематических особенностей она употребляется во много раз чаще, так что в отрывке из 100 графем может оказаться сразу несколько Q. Поэтому ее частоты в рамках одного произведения достаточно близки к теоретическим, а в рамках сразу нескольких произведений резко отличаются от них.

Если иметь в виду результаты обоих этапов эксперимента, то можно утверждать, что:

– большинство немецких графем характеризуются стабильностью своих частот;

– из этого правила выпадают только три графемы, а именно, F, K и Q, частоты которых нестабильны либо в рамках одного произведения (F, K) либо внутри массива текстов, составленного из нескольких произведений (Q).

В свете этих результатов можно рассмотреть высказанное автором предположение, что шансы на стабильность частот возрастают по мере того, как уменьшается зависимость употребления языковых единиц от воли человека (см., Гвоздович Б.Н., 1975). Само собой разумеется, что для окончательного ответа на этот вопрос необходимо иметь соответствующие данные по многим языкам. Но если исходить из того, что графемы в указанном смысле более независимы, чем фонемы, то выводы настоящего эксперимента говорят в пользу сформулированного выше предположения.

На самом деле, исследования показали, что, например, из

[+] Критическое значение χ^2 равно 576,49 при 500 степенях свободы (см. Оуэн О.Б., 1973, с. 55).

42 польских фонем стабильностью частот отличаются 23, т.е. чуть больше половины, в то время как немецкие графемы со стабильными частотами составляют почти девять десятых из общего числа (23 из 26).

ЛИТЕРАТУРА

Гвоздович Б.Н. Однородность текстов относительно частот немецких графем. – В кн.: Вопросы терминологии и лингвистической статистики. Воронеж, 1975.
Гусева Е.К. К вопросу о количественном подходе к анализу структурных особенностей буквенных знаков. – В кн.: Проблемы грамматического моделирования. М., 1973.
Оуэн О.Б. Статистические таблицы. М., 1973.
Урбах В.Ю. Биометрические методы. М., 1964.
Wolf Ch. Der geteilte Himmel. Halle (Saale), 1963.

DIE STABILITÄT DER HÄUFIGKEITEN DER DEUTSCHEN GRAPHEME

Boris Gwosdowitsch

R e s ü m e e

Der Artikel behandelt die Ergebnisse einer Untersuchung der Stabilität der Häufigkeiten der deutschen Grapheme. Die Untersuchung hat ergeben, daß die meisten deutschen Grapheme stabile Häufigkeiten besitzen. Die Grapheme F, K und Q haben dagegen unstabile Häufigkeiten – die beiden ersteren im Rahmen eines Werkes, das letztere – beim Übergang von einem Werk zu einem anderen.

Der Verfasser nimmt an, daß die sprachlichen Einheiten desto mehr Chancen auf stabile Häufigkeiten haben, je weniger ihr Gebrauch vom Willen des Menschen abhängt.

ДРЕВНЕСЛАВЯНСКИЙ ЯЗЫК И ЕГО ТИПЫ ПО ЛИНГВОСТАТИСТИЧЕСКИМ ДАННЫМ

А.С. Герд

Древнеславянский язык – общий письменный язык славян в XI–XVI веках. В свою очередь он делился на ряд подтипов в зависимости от хронологии, жанра, ареала, традиций центров письменности. Выявление и установление типов древнеславянского языка – одна из актуальных задач славянской филологии.

В 1974, 1977 годах были опубликованы две книги единого исследования ленинградских филологов – Именное склонение в славянских языках (Герд А.С. и др., 1974; 1977). В этих книгах представлены результаты обследования 104 славянских источников XI–XVI веков с общим объемом выборки свыше I 200 000 словоупотреблений при выборке в 12 000 словоупотреблений на отдельный текст (локальный центр). Для каждого текста (автора) в этих монографиях приводится полный список флексий и их частот. Настоящая статья построена на материале этих книг, откуда почерпнуты все статистические данные, там же смотри и список источников. Цель статьи – выявление разных типов древнеславянских текстов.

Выбор непротиворечивых, нейтральных критериев для типологического сравнения древних текстов – дело нелегкое. С одной стороны, типологические параметры в истории языка не должны быть чересчур оторваны от текста; они должны служить выяснению нерешенных филологических проблем в области исторического языкознания. С другой стороны, однако, эти параметры не должны зависеть от характера отдельных текстов, от их жанра, содержания, локальной приуроченности и т.д. Так, например, статистические показатели по категории падежа находятся в сильной зависимости от категорий рода, числа, от жанра; категория рода тесно связана с лексикой, а лексика, как известно, требует больших выборок; многие типы старых именных основ низкочастотны в памятниках, а флексии некоторых падежей (род пад. о – основ, м.р. флексия – у/ю; дат. пад. о – основ, м.р. флексия – ови/еви), обнаруживают яркую локальную приуроченность, независимо от их статистики.

Наконец, те или иные выводы сравнительно-типологического характера убедительны только в том случае, если они основаны на достаточно репрезентативном числе разных источ-

ников и на фактах, явлениях статистически сильных.

Настоящая статья основана на анализе 98 источников XУ-XУI веков, а при рассмотрении отдельных явлений, параметров опирается только на факты, которые в той или иной степени статистически представлены во всех текстах.

В то же время именно морфология - самое устойчивое ядро системы любого языка, - наиболее благоприятный полигон для исследований в области квантитативной типологии текстов.

Рассмотрим, во-первых, как группируются тексты XУ-XУI веков на основании сравнения данных по статистике флексий статистически сильных падежей в единственном числе. Представим с этой целью соответствующий материал в виде частотных списков флексий по убыванию их абсолютных частот.+

Ниже в списке в первой колонке слева ранг памятника (r_j), далее следует показатель абсолютной частоты на I2 000 словоупотреблений, затем - ареал текста, век, тип текстов (жанр), в скобках название конкретного памятника, имя автора, если вся выборка (I2 000) относится к этому тексту или к этому автору.

Род. падеж о - основ м.р. флексия -а/-я/-Ø

I	I00I	-	Валахия, I5, дел.
2	66I	-	Москва, I5, лет.-хрон.
3	62I	-	Брест, I6, дел.
4	588	-	Хорватия, I6, дел.
5	530	-	Сербия, I5, дел.
6	437	-	Хорватия, I5, дел.
7	437	-	Юго-зап. Русь, I6, дел.
8	434	-	Псков, I6, лет.-хрон.
9	42I	-	Рязань, I6, дел.
I0	4I7	-	Москва, I6, дел.
II	40I	-	Москва, I6, лет.-хрон.
I2	365	-	Москва, I5, пов. (Задонщина)
I3	328	-	Псков, I5, лет.-хрон.
I4	300	-	Острог, I6, пов. (Василий)
I5	298	-	Дубровник, I5, дел.
I6	295	-	Босния, I5, дел.
I7	285	-	Тверь, I6, дел.
I8	285	-	Вильно, I6, дел.
I9	266	-	Москва, I5, конф.-пов. (Иосиф Волоцкий)
20	254	-	Русь, I6, хожд.
2I	243	-	Псков, I6, пов. (Пов. о Батории)
22	236	-	Новгород, I5, пов.

+ Здесь и ниже в списках данных названия ареалов, жанров, памятников, авторов, сохранены в том же виде, как и в книгах Именное склонение в славянских языках XI-XIУ веков и Именное склонение в славянских языках XУ-XУI веков (Герд А.С. и др., 1974; 1977).

23 218 - Москва, I6, пов. (Степенная книга)
24 217 - Москва, I6, пов.
25 208 - Зап. Русь, I5, лет.-хрон.
26 208 - Зап. Русь, I6, пов. (Иван Вишенский)
27 205 - Сев. Двина, I5, дел.
28 202 - Москва, I6, дел. (Судебник)
29 200 - Москва, I6, конф.-пов.
30 200 - Болгария, I5, конф.-пов. (Владислав Граматик)
31 200 - Русь, I5, хожд.
32 200 - Валахия, I6, конф.-пов.
33 I80 - Польша, I6, дел.
34 176 - Москва, I5, конф.-пов. (Пахомий Логофет)
35 I75 - Сербия, I5, пов.
36 I64 - Болгария, I6, конф.-пов. (Матей Граматик)
37 160 - Москва, I5, конф.-пов. (Епифаний Премудрый)
38 160 - Псков, I5, дел.
39 I52 - Москва, I5, дел.
40 I47 - Валахия, I5, конф.-пов.
41 146 - Юго-зап. Русь, I6, конф. (Пересопницкое евангелие)
42 I43 - Хорватия, I6, дел. (Полички статут)
43 I40 - Чехия, I5, дел.
44 I35 - Польша, I6, пов.
45 I35 - Польша, I6, пов. (Петр Скарга)
46 I30 - Болгария, I5, конф.-пов. (Константин Костенечски)
47 I22 - Зап. Русь, I6, конф.-пов. (Евангелие Тяпинского)
48 II4 - Польша, I5, конф.-пов.
49 II3 - Москва, I6, дел. (Домострой)
50 I05 - Москва, I6, конф.-пов. (Четии-Минеи)
5I I00 - Юго-зап. Русь, I5, дел.
52 99 - Болгария, I5, конф.-пов. (Димитр Кантакузин)
53 97 - Зап. Русь, I5, дел.
54 96 - Чехия, I5, конф.-пов. (Ян Гус)
55 88 - Чехия, I5, хожд.
56 85 - Польша, I6, конф.-пов.
57 78 - Чехия, I6, пов.
58 6I - Словения, I6, конф.-пов. (Юрий Далматин)
59 5I - Чехия, пов. (Троянская история)
60 49 - Словения, I6, конф.-пов. (Примус Трубер)

Дат. падеж о - основ
м.р. - флексия - у/ю/Ѫ

I 338 - Тверь, I6, дел.
2 335 - Москва, I6, лет.-хрон.
3 3II - Валахия, I5, дел.
4 308 - Рязань, I6, дел.
5 288 - Москва, I6, дел. (Судебник)
6 282 - Москва, I5, лет.-хрон.
7 277 - Босния, I5, дел.
8 269 - Зап. Русь, I5, лет.-хрон.
9 250 - Сербия, I5, дел.
I0 224 - Русь, I6, хожд.
II 223 - Дубровник, I5, дел.
I2 220 - Москва, I5, пов. (Задонщина)
I3 2I0 - Псков, I5, лет.-хрон.
I4 200 - Хорватия, I6, дел.
I5 I9I - Москва, I6, дел.
I6 I8I - Псков, I6, пов. (Пов. о Батории)
I7 I65 - Юго-зап. Русь, I6, дел.

I8	I6I	-	**Москва**, I5, конф.-пов. (Иосиф Волоцкий)
I9	I59	-	**Москва**, I6, дел. (Домострой)
20	I59	-	**Москва**, I5, конф.-пов. (Пахомий Логофет)
2I	I57	-	**Болгария**, I6, конф.-пов. (Матей Граматик)
22	I52	-	**Болгария**, I5, конф.-пов. (Владислав Граматик)
23	I50	-	**Москва**, I6, пов. (Степенная книга)
24	I47	-	**Хорватия**, I5, дел.
25	I45	-	**Псков**, I5, дел.
26	I34	-	**Псков**, I6, лет.-хрон.
27	I33	-	**Валахия**, I6, конф.-пов.
28	I23	-	**Валахия**, I5, конф.-пов.
29	I23	-	**Москва**, I6, конф.-пов. (Четии-Минеи)
30	I22	-	**Новгород**, I5, пов.
3I	II8	-	**Болгария**, I5, конф.-пов. (Константин Констеченски)
32	II7	-	**Москва**, I6, пов.
33	I09	-	**Сербия**, I5, пов.
34	I03	-	**Острог**, I6, пов. (Василий)
35	98	-	**Вильно**, I6, дел.
36	88	-	**Болгария**, I5, конф.-пов. (Димитр Кантакузин)
37	83	-	**Брест**, I6, дел.
38	77	-	**Москва**, I5, дел.
39	75	-	**Чехия**, I6, дел.
40	72	-	**Русь**, I5, хожд.
41	72	-	**Зап. Русь**, I5, дел.
42	7I	-	**Хорватия**, I6, дел. (Полически статут)
43	68	-	**Юго-зап. Русь**, I5, дел.
44	66	-	**Чехия**, I5, дел.
45	64	-	**Москва**, I5, конф.-пов. (Епифаний Премудрый)
46	63	-	**Зап. Русь**, I6, конф.-пов. (Евангелие Тяпинского)
47	63	-	**Юго-зап. Русь**, I6, пов. (Иван Вишенский)
48	56	-	**Чехия**, I5, конф.-пов. (Ян Гус)
49	56	-	**Чехия**, I6, пов.
50	56	-	**Словения**, I6, конф.-пов. (Юрий Далматин)
5I	55	-	**Польша**, I5, конф.-пов.
52	52	-	**Юго-зап. Русь**, I6, конф.-пов. (Пересопницкое евангелие)
53	38	-	**Польша**, I6, конф.-пов.
54	35	-	**Чехия**, I5, хожд.
55	34	-	**Польша**, I6, пов.
56	28	-	**Чехия**, I5, пов. (Троянская история)
57	28	-	**Сев. Двина**, I5, дел.
58	27	-	**Словения**, I6, конф.-пов. (Примус Трубер)
59	2I	-	**Польша**, I6, дел.
60	I4	-	**Польша**, I6, пов. (Петр Скарга)

Тв. пад. о - основ

м.р. флексия - ъмъ/ъмь

I	2I0	-	**Рязань**, I6, дел.
2	20I	-	**Босния**, I5, дел.
3	I38	-	**Хорватия**, I5, дел.
4	I3I	-	**Брест**, I6, дел.
5	I27	-	**Валахия**, I5, дел.
6	I26	-	**Москва**, I6, лет.-хрон.
7	II6	-	**Зап. Русь**, I6, пов. (Иван Вишенский)
8	II4	-	**Псков**, I5, лет.-хрон.
9	I09	-	**Псков**, I6, лет.-хрон.
I0	I09	-	**Вильно**, I6, дел.
II	I08	-	**Хорватия**, I6, дел.
I2	I08	-	**Москва**, I6, дел.

13 108 - Москва, 15, лет.-хрон.
14 100 - Псков, 16, пов. (Пов. о Батории)
15 95 - Болгария, 15, конф.-пов. (Владислав Граматик)
16 95 - Чехия, 16, пов.
17 94 - Юго-зап. Русь, 16, дел.
18 93 - Сербия, 15, пов.
19 91 - Москва, 15, пов. (Задонщина)
20 91 - Москва, 16, конф.-пов. (Четии-Минеи)
21 90 - Новгород, 15, пов.
22 85 - Москва, 16, пов. (Степенная книга)
23 82 - Болгария, 15, конф.-пов. (Константин Костенечски)
24 82 - Русь, 16, хожд.
25 80 - Болгария, 16, конф.-пов. (Матей Граматик)
26 79 - Чехия, 15, дел.
27 77 - Зап. Русь, 15, лет.-хрон.
28 76 - Москва, 16, дел.
29 75 - Юго-зап. Русь, 16, конф.-пов. (Пересопницкое евангелие)
30 74 - Валахия, 16, конф.-пов.
31 72 - Чехия, 15, конф.-пов. (Ян Гус)
32 69 - Москва, 15, конф.-пов.
33 68 - Тверь, 16, дел.
34 66 - Москва, 15, конф.-пов. (Пахомий Логофет)
35 66 - Острог, 16, пов. (Василий)
36 65 - Польша, 16, пов.
37 64 - Москва, 15, конф.-пов. (Епифаний Премудрый)
38 60 - Москва, 16, дел. (Домострой)
39 56 - Чехия, 15, хожд.
40 54 - Москва, 16, пов.
41 51 - Валахия, 15, конф.-пов.
42 51 - Сербия, 15, дел.
43 50 - Псков, 15, дел.
44 49 - Чехия, 15, пов. (Троянская история)
45 48 - Москва, 15, дел.
46 48 - Сев. Двина, 15, дел.
47 48 - Польша, 16, пов. (Петр Скарга)
48 47 - Чехия, 16, дел.
49 46 - Русь, 15, хожд.
50 42 - Болгария, 15, конф.-пов. (Димитр Кантакузин)
51 42 - Хорватия, 16, дел.
52 37 - Зап. Русь, 15, дел.
53 36 - Польша, 15, конф.-пов.
54 35 - Польша, 16, конф.-пов.
55 33 - Юго-зап. Русь, 15, дел.
56 32 - Дубровник, 15, дел.
57 28 - Зап. Русь, 15, конф.-пов. (Евангелие Тяпинского)
58 27 - Словения, 16, конф.-пов. (Юрий Далматин)
59 26 - Польша, 16, дел.
60 23 - Словения, 16, конф.-пов. (Примус Трубер)

Местн. падеж о - основ.

м.р. - флексия ѣ/е

1 197 - Тверь, 16, дел.
2 121 - Москва, 16, дел. (Судебник)
3 101 - Москва, 16, дел.
4 93 - Псков, 15, лет.-хрон.
5 87 - Юго-зап. Русь, 16, дел.
6 85 - Москва, 15, лет.-хрон.
7 78 - Псков, 16, лет.-хрон.

8	78	-	Болгария, I6, конф.-пов. (Матей Граматик)
9	77	-	Острог, I6, пов. (Василий)
I0	70	-	Русь, I6, хожд.
II	68	-	Москва, I6, пов. (Степенная книга)
I2	6I	-	Валахия, I6, конф.-пов.
I3	60	-	Москва, I6, лет.-хрон.
I4	59	-	Русь, I5, хожд.
I5	58	-	Болгария, I5, конф.-пов. (Владислав Граматик)
I6	57	-	Болгария, I5, конф.-пов. (Константин Костенечски)
I7	57	-	Польша, I6, пов. (Петр Скарга)
I8	56	-	Польша, I6, конф.-пов.
I9	56	-	Вильно, I6, дел.
20	55	-	Рязань, I6, дел.
2I	55	-	Москва, I5, конф.-пов. (Иосиф Волоцкий)
22	55	-	Москва, I6, пов.
23	53	-	Москва, I5, конф.-пов. (Пахомий Логофет)
24	50	-	Псков, I5, дел.
25	50	-	Зап. Русь, I5, лет.-хрон.
26	49	-	Москва, I6, дел. (Домострой)
27	49	-	Москва, I5, дел.
28	49	-	Москва, I6, конф.-пов. (Четии-Минеи)
29	48	-	Чехия, I5, хожд.
30	48	-	Брест, I6, дел.
3I	43	-	Валахия, I5, дел.
32	43	-	Сев. Двина, I5, дел.
33	4I	-	Псков, I6, пов. (пов. о Батории)
34	40	-	Юго-зап. Русь, I6, конф.-пов. (Пересопницкое евангелие)
35	38	-	Москва, I5, конф.-пов. (Епифаний Премудрый)
36	38	-	Зап. Русь, I6, пов. (Иван Вишенский)
37	34	-	Польша, I6, пов.
38	3I	-	Сербия, I5, пов.
39	30	-	Валахия, I5, конф.-пов.
40	28	-	Болгария, I5, конф.-пов. (Димитр Кантакузин)
4I	27	-	Польша, I5, конф.-пов.
42	26	-	Хорватия, I5, дел.
43	25	-	Чехия, I5, конф.-пов. (Ян Гус)
44	25	-	Юго-зап. Русь, I5, дел.
45	23	-	Зап. Русь, I6, конф.-пов. (Евангелие Тяпинского)
46	20	-	Чехия, I5, дел.
47	20	-	Чехия, I6, пов.
48	I8	-	Москва, I5, пов. (Задонщина)
49	I8	-	Польша, I6, дел.
50	I7	-	Зап. Русь, I5, дел.
5I	I5	-	Хорватия, I6, дел.
52	7	-	Новгород, I5, пов.
53	7	-	Чехия, I6, дел.
54	6	-	Чехия, I5, пов.
55	6	-	Сербия, I5, дел.
56	2	-	Хорватия, I6, дел. (Полически статут)
57	2	-	Словения, I6, конф.-пов. (Примус Трубер)
58	I	-	Босния, I5, дел.
59	I	-	Дубровник, I5, дел.
60	I	-	Словения, I6, конф.-пов. (Юрий Далматин)

Частотные списки по отдельным флексиям свидетельствуют о том, что независимо от хронологии и ареала выделяются два основных типа древнеславянского языка.

I. Язык деловых и летописно-хроникальных текстов (в

среднем это зона $\tau_i \approx$ I-19; I-25)

2. Язык текстов конфессионально-повествовательных, повествовательных и хождений ($\tau_i \approx$ 20-60).

Параллельно в рамках общего второго типа выделяется подтип языка западнославянских текстов, объединяющий в себе тексты из Западной и Юго-западной Руси, Польши и Чехии.[+] ($\tau_i \approx$ 30-60)

Во всех случаях обнаруживаются переходные зоны, включающие тексты соседних типов.

Во вторых, рассмотрим, как группируются те же тексты XУ-XУI веков по такому статистическому параметру, как математическое ожидание падежа (М). (Исходные данные см. Герд А.С. и др., I977.)

I	404,38	-	I5,	Валахия, дел.[++]
2	356,I2	-	I6,	Тверь, дел.
3	347,37	-	I6,	Москва, дел.
4	3I0,45	-	I6,	Хорватия, дел.
5	224,8	-	I5,	Хорватия, дел.
6	222,74	-	I6,	Брест, дел.
7	208,97	-	I6,	Рязань, дел.
8	202,37	-	I6,	Псков, лет.-хрон.
9	I97,76	-	I5,	Москва, лет.-хрон.
I0	I90,75	-	I6,	Москва, лет.-хрон.
II	I82,II	-	I6,	Юго-зап. Русь, дел.
I2	I53,9	-	I6,	Москва, конф.-пов. (Четии-Минеи)
I3	I3I,9	-	I5,	Босния, дел.
I4	I28,32	-	I5,	Псков, лет.-хрон.
I5	I27,94	-	I6,	Вильно, дел.
I6	I27,9	-	I6,	Валахия, конф.-пов.
I7	I25,92	-	I5,	Сербия, дел.
I8	I20,06	-	I6,	Болгария, конф.-пов. (Матей Граматик)
I9	II9,55	-	I6,	Москва, дел. (Судебник)
20	II9,44	-	I5,	Сербия, пов.
2I	I09,99	-	I5,	Болгария, конф.-пов. (Владислав Граматик)
22	I06,69	-	I5,	Западная Русь, лет.-хрон.
23	I0I,35	-	I6,	Острог, пов. (Василий)
24	I0I,I8	-	I6,	Москва, пов.
25	99,77	-	I6,	Москва, конф.-пов. (Степенная книга)
26	97,02	-	I6,	Псков, пов. (Пов. о Батории)
27	95,76	-	I5,	Русь, хожд.
28	93,5	-	I5,	Москва, пов. (Задонщина)
29	9I,52	-	I6,	Пересопница, конф.-пов.
30	88,2	-	I6,	Русь, хожд.
3I	86,77	-	I5,	Чехия, конф.-пов. (Ян Гус)

[+] Эти типы языков выделяются не по принципу исключительного, а преимущественного преобладания текстов данного типа в той или иной зоне.

[++] Здесь в первой колонке - ранг (τ_i), во второй -математическое ожидание (М), в третьей - век, в четвертой - место создания текста, в пятой - тип текста.

32	85,27	-	I5,	Новгород, пов.
33	82,25	-	I5,	Москва, конф.-пов. (Иосиф Волоцкий)
34	8I,78	-	I5,	Дубровник, дел.
35	8I,46	-	I6,	Москва, дел. (Домострой)
36	74,75	-	I6,	Юго-зап. Русь, пов. (Иван Вишенский)
37	68,08	-	I5,	Болгария, конф.-пов. (Константин Костенечски)
38	66,87	-	I5,	Северная Двина, дел.
39	63,25	-	I5,	Чехия, пов. (Троянская история)
40	63,I5	-	I6,	Западная Русь, конф.-пов. (Евангелие Тяпинского)
4I	62,I5	-	I5,	Чехия, хожд.
42	6I,34	-	I5,	Чехия, пов.
43	59,77	-	I5,	Москва, конф.-пов. (Пахомий Логофет)
44	56,96	-	I5,	Чехия, дел.
45	56,34	-	I5,	Валахия, конф.-пов.
46	5I,65	-	I5,	Польша, конф.-пов.
47	5I,5I	-	I5,	Псков, дел.
48	50,24	-	I6,	Польша, пов.
49	50,I	-	I6,	Чехия, дел.
50	48,6	-	I5,	Болгария, конф.-пов. (Димитр Кантакузин)
5I	48,3	-	I5,	Москва, дел.
52	47,56	-	I5,	Москва, конф.-пов. (Епифаний Премудрый)
53	42,32	-	I6,	Словения, конф.-пов. (Юрий Далматин)
54	38,2I	-	I6,	Польша, конф.-пов.
55	37,6I	-	I6,	Польша, дел.
56	3I,38	-	I6,	Польша, пов. (Петр Скарга)
57	27,86	-	I5,	Юго-зап. Русь, дел.
58	24,2I	-	I5,	Зап. Русь, дел.

По приведенным данным, независимо от хронологии и ареала выделяются два основных типа языка: I. Язык деловых и летописно-хроникальных текстов ($\tau_i \approx$ I-I7); 2. Язык текстов конфессионально-повествовательных, повествовательных и хождений ($\tau_i \approx$ I8-58). Отдельно, как параллельно сосуществующие можно выделить следующие ареально-жанровые подтипы древнеславянского языка - тип восточно-славянских и южнославянских деловых и летописных текстов ($\tau_i \approx$ I-I7), - тип южновосточнославянских - преимущественно конфессионально-повествовательных и повествовательных текстов ($\tau_i \approx$ I8-40), - тип западнославянских текстов, исключая тексты южно-восточнославянские ($\tau_i \approx$ 39-56); - язык чешских текстов с М. в интервале 63-56 ($\tau_i \approx$ 40-44); - язык польских и западнорусских текстов с М. в интервале 5I-24 ($\tau_i \approx$ 46-58).

Особо отметим близость языка текстов Москвы, Пскова и Западной и Юго-западной Руси (№№ 7,8,9,I0,II,I2,I4,I5,24,25, 26,28,29) и отдельно текстов Пскова к Западной Руси и Польше (τ_i = I4,I5,46,47).

Ареально-жанровые подтипы существуют параллельно; и здесь везде наблюдаются переходные звенья, зоны (τ_i = II, I6,I9,22 и др.). Ряд центров (τ_i = 34,38,47,5I) выделяются

хронологически (дел., 15 в.)

Наконец, один из важных критериев типологического сопоставления древнеславянских памятников письменности на статистической основе - степень насыщенности текста флексиями. Показателем такой насыщенности является общая суммарная накопленная частота всех флексий, встретившихся в выборке из памятника. Рассмотрим этот последний параметр не только по памятникам XУ-XУI веков, но по данным 98-и источников с XI-XУI века, привлеченных в названных выше книгах.

Приведем списки славянских текстов и памятников XI-XУI веков по степени их насыщенности флексиями. Ниже в первой колонке слева представлен ранговый номер (r_i), во второй - абсолютная накопленная частота всех флексий в тексте, в третьей - относительная накопленная частота к выборке в 12 000 словоупотреблений, в четвертой - место создания текста, текстов, в пятой - их хронология, в шестой - тип текста, а в скобках, в ряде случаев, конкретные памятники, авторы.

I	5132	-	0,42	Тверь,	I6,	дел.
2	4998	-	0,41	Москва,	I6,	дел.
3	4764	-	0,39	Валахия,	I5,	дел.
4	4487	-	0,37	Хорватия,	I6,	дел.
5	4336	-	0,36	Псков,	I5,	лет.-хрон.
6	4216	-	0,35	Русь,	I4,	лет.-хрон.
7	4160	-	0,34	Москва,	I6,	лет.-хрон.
8	4144	-	0,34	Москва,	I5,	лет.-хрон.
9	3981	-	0,33	Псков,	I6,	лет.-хрон.
I0	3386	-	0,33	Брест,	I6,	дел.
II	3829	-	0,3I	Новгород,	I3,	лет.-хрон.
I2	3774	-	0,3I	ЮЗР,	I6,	дел.
I3	3687	-	0,30	Рязань,	I6,	дел.
I4	36I4	-	0,30	Хорватия,	I5,	дел.
I5	3477	-	0,28	Русь,	I4,	лет.-хрон.
I6	3416	-	0,28	Москва,	I6,	дел. (Судебник)
I7	34I3	-	0,28	Валахия,	I6,	конф.-пов.
I8	3393	-	0,28	Вильно,	I6,	дел.
I9	3326	-	0,27	Москва,	I6,	конф.-пов. (Четии-Минеи)
20	3200	-	0,26	Сербия,	I5,	пов.
21	3206	-	0,26	Болгария,	I4,	канон. (Синодик царя Борила)
22	3197	-	0,26	Босния,	I5,	дел.
23	3151	-	0,26	Русь,	I3,	конф.-пов. (Житие Саввы)
24	3190	-	0,26	Болгария,	I6,	конф.-пов. (М.Граматик)
25	3141	-	0,26	Псков,	I6,	пов. (Пов. о Батории)
26	3090	-	0,25	Русь,	I4,	конф.-пов. (Заветы I2 патриархов)
27	3059	-	0,25	Словения,	I6,	конф.-пов. (Пр. Трубер)
28	3039	-	0,25	Болгария,	I5,	конф.-пов. (Вл. Грамматик)
29	2971	-	0,24	Москва,	I6,	пов.
30	2960	-	0,24	Москва,	I5,	пов. (Задонщина)
31	2947	-	0,24	Русь,	I4,	канон. (Мерило праведное)
32	2947	-	0,24	Сербия,	I5,	дел.

33	2912	-	0,24	Русь,	I6,	хожд.
34	2907	-	0,24	Москва,	I6,	пов.
35	2873	-	0,23	Острог,	I6,	пов. (Василий)
36	2855	-	0,23	Новгород,	II,	конф. (Минеи)
37	2847	-	0,23	Москва,	I5,	конф.-пов. (И. Волоцкий)
38	2830	-	0,23	Москва,	I6,	дел. (Домострой)
39	2814	-	0,23	Русь,	I3,	конф. (Симоновская псалтырь)
40	2786	-	0,23	Пересопница,	I6,	конф.-пов.
41	2721	-	0,22	Русь,	I4,	конф. (Чуд.нов. завет)
42	2649	-	0,22	Чехия,	I6,	пов.
43	2610	-	0,21	Сербия,	I4,	конф. (Шишатовацкий апостол)
44	2605	-	0,21	Чехия,	I5,	конф.-пов. (Ян Гус)
45	2558	-	0,2I	Русь,	I2,	канон. (Ефрем. кормчая)
46	2549	-	0,2I	Чехия,	I5,	пов. (Троянская история)
47	2513	-	0,20	Дубровник,	I5,	дел.
48	2474	-	0,20	Болгария,	I5,	конф.-пов. (К. Костенечски)
49	2447	-	0,20	ЮЗР,	I6,	пов. (И. Вишенский)
50	2428	-	0,20	Русь,	I3,	конф.-пов. (Житие Нифонта)
5I	242I	-	0,20	Чехия,	I5,	дел.
52	2418	-	0,20	Русь,	I2,	конф.-пов. (Слова И. Златоуста)
53	2395	-	0,I9	Русь,	I3,	конф. (Толк. апостол)
54	2381	-	0,I9	Сев. Двина	I5,	дел.
55	2362	-	0,I9	Русь,	I2,	конф.-пов. (Сказание о Борисе и Глебе)
56	2350	-	0,I9	Болгария,	I3,	конф. (Добрейшево, ев.)
57	2326	-	0,I9	Сербия,	I2,	конф. (Мирославово, ев.)
58	2324	-	0,I9	Зап. Русь	I5,	лет.-хрон.
59	2319	-	0,I9	Чехия,	I6,	дел.
60	2315	-	0,I9	Сербия,	I4,	канон. (Синодик)
6I	2314	-	0,I9	Зап. Русь,	I6,	конф. (Ев. Тяпинского)
62	2293	-	0,I9	Русь,	I2,	конф. (Галицкое ев.)
63	2268	-	0,I8	Русь,	I2,	конф.-пов.(Житие Феод. Печерского)
64	2229	-	0,I8	Русь,	I3,	канон. (Новг. кормчая)
65	2226	-	0,I8	Русь,	II,	конф.-пов. (Изборник, 1073 г.)
66	22I6	-	0,I8	Москва,	I5,	конф.-пов. (Пахомий Логофет)
67	2203	-	0,I8	Русь,	II,	конф.-пов. (Изборник, 1076 г.)
68	2I96	-	0,I8	Валахия,	I5,	конф.-пов. (Еп. Премудрый)
69	2I86	-	0,I8	Москва,	I5,	конф.-пов.
70	2I8I	-	0,I8	Болгария,	I4,	лет.-хрон. (Манасиева хроника)
7I	2I59	-	0,I7	Польша,	I5,	конф.-пов.
72	203I	-	0,I6	Москва,	I5,	дел.
73	2025	-	0,I6	Русь,	II,	конф. (Остромирово ев.)
74	195I	-	0,I6	Новгород,	I5,	пов.
75	1947	-	0,I6	Польша,	I6,	пов.
76	1904	-	0,I5	Болгария,	I4,	конф.-пов. (Евф. Тырновский)
77	1899	-	0,I5	Болгария,	I2,	конф. (Охридский апостол)

78	I846	-	0,I5	Болгария,	I3,	конф. (Болонская пс.)
79	I825	-	0,I5	Болгария,	I5,	конф.-пов. (Дм. Кантакузин)
80	I824	-	0,I5	Болгария,	I2,	конф. (Погодинская пс.)
81	I807	-	0,I5	Сербия,	I4,	конф. (Бухарестская пс.)
82	I804	-	0,I5	Болгария,	II,	конф. (Синайская пс.)
83	I795	-	0,I4	Болгария,	I4,	конф. (Софийская пс.)
84	I748	-	0,I4	Псков,	I5,	дел.
85	I738	-	0,I4	Сербия,	I3,	конф. (Вуканово ев.)
86	I72I	-	0,I4	Польша,	I6,	конф.-пов.
87	I635	-	0,I3	Польша,	I6,	пов. (П. Скарга)
88	I607	-	0,I3	Болгария,	II,	конф. (Саввина кн.)
89	I600	-	0,I3	Польша,	I6,	дел.
90	I594	-	0,I3	Хорватия,	I6,	дел. (Полически статут)
9I	I583	-	0,I3	Русь,	II,	конф. (Чудовская пс.)
92	I525	-	0,I2	Болгария;	I2,	конф. (Слепченский апостол)
93	I449	-	0,I2	Зап. Русь,	I5,	дел.
94	I439	-	0,II	Болгария,	I2,	конф. (Григоровичев паремейник)
95	I427	-	0,II	Болгария,	II,	конф. (Зографское ев.)
96	I405	-	0,II	ИЗР,	I5,	дел.
97	I339	-	0,II	Болгария,	II,	конф. (Марминское ев.)
98	I274	-	0,I0	Русь,	II,	конф.-пов. (Синайский патерик)

Во-первых, четко выделяется зона (I-I8) - зона деловых, летописных и хроникальных текстов разных ареалов. Во-вторых - зона (I8-72) - зона повествовательных и конфессионально-повествовательных текстов разных веков и ареалов. Здесь концентрируются прежде всего повести, жития и близкие к ним тексты хождений. Нахождение здесь Домостроя свидетельствует как раз о том, что в определенных отношениях этот памятник ближе к текстам повествовательным. В конце этой зоны обнаруживаем довольно широкое переходное звено (56-72) к текстам собственно конфессиональным. Так здесь встречаем уже и ряд евангелий (Добрейшево, Мирославово, Галицкое, Тяпинского). Здесь же, во второй зоне отдельно по ареальному и хронологическому признаку выделяется самостоятельная микрозона П-1 - южнославянских и чешских деловых текстов XУ в. (Босния, Сербия, Дубровник, Чехия). В этой связи наличие в конце первой зоны хорватских деловых текстов XУ в. представляет скорее переход именно к микрозоне П-1. В-третьих, зона в интервале 73-98. Эта зона внутренне членится на ряд микрозон: а) микрозона Ш-1 - зона южно-восточно-славянских собственно конфессиональных текстов XI-XIУ веков (Болгария, Сербия, Русь); б) микрозона Ш-2 - зона западнославянских повествовательных и конфессионально-повествовательных сочинений XУ-XУI веков; в) микрозона Ш-3 - зона русских и западнорусских деловых документов XУ в.

Своеобразие данной зоны в целом в том, что она выбрала в себя тексты разных эпох, школ и жанров. В то же время нетрудно заметить, что при внешнем формальном совпадении в одной зоне эти микрозоны содержательно отражают и разные эпохи, и разные ареалы и жанры. Все приведенные данные вновь свидетельствуют о том, что определяющим параметром при характеристике языка памятника древнеславянской письменности является не наличие - отсутствие флексий, не хронология, а именно жанр, тип текста. В XI-XVI веках в рамках одного общего жанра (типа текстов) ни время, ни место создания памятника, ни этнос не определяют лингвистический тип языка. Славянские тексты XI-XVI веков прежде всего различались по языку жанра, традиций школы. В этой связи следует лишний раз подчеркнуть точность "традиционных" интуитивных терминов "язык деловых памятников", "язык повести", "язык конфессиональных текстов" и т.п.+

Итак, избранный морфолого-статистический критерий позволяет выделить следующие общие и частные типы древнеславянских текстов:

общие типы: 1) тип деловых и летописно-хроникальных текстов;

2) тип повествовательных и конфессионально-повествовательных текстов.

Частные типы. Среди текстов первого типа выделяются: а) тип южнославянских и чешских деловых текстов XV в. б) тип русских и западнорусских деловых текстов XV в. Среди текстов второго типа отметим:

а) тип южно-восточнославянских собственно конфессиональных текстов XI-XIV веков;

б) тип западнославянских повествовательных и конфессионально-повествовательных текстов XV-XVI веков.

Общий второй тип в лингвистическом отношении, по-видимому, являет собой церковнославянский язык как таковой, представленный как восточнославянскими и южнославянскими памятниками, так и источниками западнославянскими. Тогда

+ Ср. аналогичный вывод, сделанный по данным сильных типов именного склонения о том, что все конфессиональные и конфессионально-повествовательные тексты представляют один единый тип языка. См. Герд А.С. Ареальная типология славянских текстов XIV-XVI веков. Советское славяноведение, 1982, № 5, с. 79.

для эпохи славянского средневековья выделяются два основных типа языка – летописно-деловой и церковнославянский. Эти два типа по-разному репрезентируются в реальных славянских текстах разных эпох и ареалов (ср., например, с одной стороны – тип южнославянских и чешских деловых текстов XУ в., а с другой – тип западнославянских конфессионально-повествовательных текстов XУ-XУI веков, южно-восточнославянский тип конфессиональных текстов XI-XIУ веков). Между отдельными типами, зонами нетрудно заметить отсутствие жестких границ, наличие переходных участков, памятников, но в каждой зоне четко выделяется ее ядро. И, напротив, текст которой характеризуется одним из заданных выше типов распределений и статистически принадлежит к одной из выделенных зон следует считать текстом, относящимся к данному типу древнеславянского языка. Выделенные выше типы языка (текстов) следует в дальнейшем проверить на материале глаголов, причастий, словообразования и главное – синтаксиса.

Л И Т Е Р А Т У Р А

Герд А.С., Капорулина Л.В., Колесов В.В., Рускова М.П., Черепанова О.А. Именное склонение в славянских языках XI-XIУ вв. (Лингвостатистический анализ по материалам памятников древнеславянской письменности). Л., 1974.

Герд А.С. и др. Именное склонение в славянских языках XУ-XУI вв. (Лингвостатистический анализ). Л., 1977.

OLD SLAVONIC AND ITS TYPES ON THE STATISTICAL BASIS

Alexandr S. Heard

S u m m a r y

The article deals with typological problems of Old Slavonic texts of the 11-16th centuries in terms of the noun declension. The article is based on the total sample of 1,200,000 word forms, the sample from each text being 12,000 word forms. The following general types of Old Slavonic texts of the 11-16th centuries are distinguished on the statistical basis - the type of business texts and chronicles, the type of narrations and religious narratives, and some minor local subtypes.

СПЕЦИФИКА РУССКИХ ТЕКСТОВ ПО УПОТРЕБИТЕЛЬНОСТИ В НИХ АБЗАЦЕВ С РАЗЛИЧНЫМ ПРЕДМЕТНО-ЛОГИЧЕСКИМ СОДЕРЖАНИЕМ

А.В. Зубов

Известно, что каждый текст является отражением некоторой реальной или мнимой ситуации. Однако, в связи с ограниченностью оперативной памяти человека и невозможностью полного охвата событий, человеческое сознание в процессе познания вычленяет некоторый фрагмент этих событий, некий "снятый момент", который позволяет рассмотреть события во всех или характерных особенностях, формах, связях (Гальперин, 1980, с. 517; Чесноков, 1982, с. 41). Такое членение действительности на фрагменты проводится не произвольно, а в соответствии с социально отработанными моделями, сформировавшимися в ходе длительного исторического развития человечества (Чесноков, 1982, с. 42; Сахарный, 1974, с. 41).

В виду неразрывности мышления и языка, каждому отраженному в сознании фрагменту действительности в некоторое соответствие ставятся определенные языковые формы. Одни исследователи отмечают, что разным элементам реальной действительности в мышлении соответствуют фиксированные слова (Верещагин, 1980, с. 64), другие ставят им в соответствие словосочетания, предложения и целые тексты (Денисов, 1969, с. 20). Третья группа авторов, не отрицая определенной роли указанных языковых единиц, ведущую роль в отражении фрагментов реальной действительности отдают логическому аппарату, позволяющему получать выводные знания из известных индивиду обобщенных эмпирических фактов (Кацнельсон, 1972, с. 111).

Ученые самых различных направлений отмечают, что в процессе человеческого познания наиболее важные и часто повторяющиеся отношения реальной действительности закрепляются в форме синтаксических структур. Эти структуры усваиваются нами с детства вместе со словами и звуками родного языка. Когда мы говорим, мы очень часто, не задумываясь, выбираем один из наиболее привычных для нас в данной ситуации языковых шаблонов (Адмони, 1976, с. 3; Пешковский, 1920, с. 427). Таким образом, такие шаблоны ("синтаксические модели", "структурные схемы предложений", "конструкции", "фразовые стереотипы", "синтаксические формы" и т.д. и т.п.) являются психическими реальностями, существующими в нашем

сознании вместе с единицами других уровней языка, участвующими в процессе создания текста.

Вместе с тем, "по данным современной психологии, процесс вербальной конкретизации мысли не обязательно имеет своим результатом одно предложение. Речевое действие может включать несколько предложений, каждое из которых по отношению друг к другу находится в состоянии некоторой зависимости" (Зарубина, 1968, с. 146), т.е. в языке существуют определенные структурно-смысловые модели для выражения отношений между мыслями в структурах, находящихся за пределами предложения. Как правило, цепочку таких взаимосвязанных мыслей называют высказыванием. Таким образом, высказывание "строится по законам сцепления мыслей-дискурса, где основным правилом выступают законы логического следования рассуждения, доказательства, выводы или связное описание" (Колшанский, 1976, с. 246). Основная особенность таких единиц заключается в том, что они обладают определенными и относительно устойчивыми типическими формами построения целого. Такими единицами мы уверенно пользуемся, но теоретически можем и не знать об их существовании. Они приходят к нам в процессе обучения языку и жизни, соотносятся с определенными фрагментами действительности и несут в себе не только собственный опыт автора, но и социальный опыт общества (Бахтин, 1979, с. 257-260; Stalnaker, 1972, с. 384; Глаголев, 1976, с. 106). Наличие таких синтаксических шаблонов, охватывающих несколько предложений, связано с общим положением об опережающем отражении действительности, разработанным советскими учеными Д.Н. Узнадзе и П.К. Анохиным (Анохин, 1962; Узнадзе, 1966). В соответствии с этой теорией "человек обладает способностью на основе врожденного и приобретенного опыта предвидеть развитие ситуативных обстоятельств и настраиваться на определенную реакцию с некоторым упреждением вероятностного хода событий" (Колшанский, 1983, с. 50). В процессе создания текста это положение реализуется в виде составления некоторой общей схемы развития мысли и, следовательно, её речевого оформления (Колшанский, 1983, с. 50).

Наличие речевых шаблонов-высказываний отмечается в исследованиях лингвистов, психологов и психолингвистов и специалистов по обучению языкам. Причем набор таких, шаблонов конечен, относительно невелик и специфичен для текстов одного автора или одной достаточно узкой предметной области

(Солнцев, 1977, с. 297; Бахтин, 1979, с. 259; Кестен, 1975, с. 256-264). Такие синтаксические единые объединения, включающие несколько предложений, были названы "сверхфразовыми единствами" или "сложными синтаксическими целыми". Исследования последних лет позволяют допустить, что и абзац письменного текста также является "логическим слепком" с действительности (Серкова, 1980, с. 111). Это также подтверждают и лингвистические исследования, и эксперименты психологов и психолингвистов. Так, в одной серии опытов испытуемым предъявлялись одни и те же тексты, но подготовленные по-разному: с разбивкой на авторские абзацы, без абзацев и с абзацами, не учитывающими логическую структуру текста. В итоге первые тексты воспринимались уверенно, легко и быстро. Во 2-ом и 3-ем случаях на усвоение таких текстов было потрачено гораздо больше времени, и текст усваивался плохо или совсем не усваивался (Страхова, 1971, с. 156).

В другой серии опытов художественный текст ("Описание весеннего дня"), научный текст-описание ("Воля") и научно-популярный текст ("Железо") разделялись на предложения, и испытуемым предлагалось объединить их в абзацы, а абзацы - в связный текст (Бондаренко, 1978).

Для художественного текста практически все испытуемые объединили предложения в сходные смысловые группы. Различия касались лишь порядка расположения предложений внутри смысловых групп. Примерно 1/3 испытуемых сложили текст из смысловых кусков так же, как и автор. В других же случаях наблюдалась перестановка смысловых групп (Бондаренко, 1978, с. 82-83).

Второй текст, научное описание, отличался строгой логической упорядоченностью фактов и все варианты текстов, полученных испытуемыми, были близки к оригиналу (Бондаренко, 1978, с. 83).

Научно-популярный характер третьего текста, наличие в нем нескольких предметов обсуждения привели к тому, что разные испытуемые относили одно и то же предложение в разные смысловые группы. Было получено большое разнообразие вариантов текста, но тем не менее, абзацы испытуемыми объединялись в четко прослеживаемые более крупные смысловые куски, такие как "Значение железа", "Добыча руды", "Выплавка чугуна" и т.п. (Бондаренко, 1978, с. 84).

Наконец, еще одним доказательством того, что абзац является некоторым смысловым и логическим единством, служит

эксперимент, проведенный в Мичиганском университете (Зарубина, 1971; Koen, 1968). Было показано, что у 80 % испытуемых расстановка абзацев одного и того же текста совпала. Причем не во всех случаях такая расстановка совпала с авторской.

Таким образом, приведенные факты показывают, что абзацы автором выделяются не произвольно. Они выстраиваются в тексте по определенным правилам, более строгим и стереотипным для научных текстов и более свободным и творческим в текстах художественных (Серкова, 1980, с. 111; Бондаренко, 1978, с. 89; Перекальская, 1980).

Основная особенность текста, как единого целого, заключается в том, что в нем между предложениями и их комплексами устанавливаются смысловые связи на основе отношений между объектами и явлениями действительности. При этом одни предметы или явления действительности объединяет их последовательность во времени, другие - то, что они являются составными частями или качествами одного одновременно воспринимаемого предмета, связанные пространственно или иным образом, третьи - то, что один предмет (или явление) служит причиной другого, а третий выступает следствием первого (Мильчин, 1980, с. 116).

О таких отношениях, закрепляемых языком в стабильных речевых структурах, писал еще М.В. Ломоносов. Он называл их повествованиями, описаниями, рассуждениями (Ломоносов, 1953, с. 332).

Дальнейшее развитие такой подход нашел в исследовании (Нечаева, 1974), где предпринята попытка выявить не только содержательные, но и структурные особенности текстовых образований. Отмечается, что в тексте "предложения интегрируются в типы высказывания, или типы речи (описание, повествование, рассуждение), каждый из которых имеет свое определенное функционально-смысловое значение и стабильную структурно-языковую характеристику (Нечаева, 1974, с. 228).

Различают несколько структурных разновидностей повествования, описания и рассуждения (Нечаева, 1974, с. 60-166; Васильев, 1981, с. 87-93; Гришина, 1982; Егоров, 1982; Кожин, 1982, с. 195-196). В то же время исследователи отмечают, что каждый достаточно протяженный текст одного типа может содержать отдельные "куски" текстов иного типа. Тот или иной тип текста является ведущим или преобладающим, обуславливая последовательность и структурно-логическую связь пред-

ложений (сверхфразовых единств, параграфов, глав) (Мильчин, 1980, с. 118). Наиболее ярко особенности повествования, описания и рассуждения проявляются в пределах наименьших семантико-синтаксических единиц - абзацев. Поэтому можно было бы описывать содержание текста путем задания последовательностей абзацев с определенным функционально-смысловым содержанием. Но функционально-смысловые типы представляют собой не что иное, как наиболее общие типовые способы изложения содержания и трудно будет выявить специфику в организации различных текстов, относящихся к одному функциональному стилю.

Дальнейшей конкретизацией содержания абзаца по сравнению с рассмотренным типовым способом изложения является понятие предметно-логического содержания абзаца (Гришина, 1982, с. 54-58; Ларина, 1982, с. 206-209). Например, в работе (Гришина, 1982, с. 54) под предметно-логическим содержанием описания в научном тексте понимается перечисление фактов объективной действительности (признаков, явлений, событий, взглядов, теорий), так или иначе связанных с сообщением. Однако для каждой конкретной микроситуации, описываемой абзацем, важно, какие именно факты объективной действительности перечисляются в абзаце-описании, какие из них являются основными, определяющими "костяк", основу содержания, а какие могут быть опущены или заменены другими объектами, в чем-то с ними эквивалентными.

Аналогично в определениях предметно-логического содержания абзаца-рассуждения и абзаца-повествования (Гришина, 1982, с. 55-56) не дифференцируется, чьи (какие) мысли высказываются в абзаце, кто или что выполняет какие-то действия или находится в том или ином состоянии. Такая детализация содержания требует предварительного установления в тексте определенной иерархии лексических единиц, отражающих денотаты. Не рассматривая детально этот аспект содержания (подробнее см., например,: Лингвистические вопросы..., 1983, с. 123; Смирнов, 1966), причем, что в тексте можно выделить две группы опорных слов (или "смысловых вех", "важных слов", "дескрипторов" и т.п.).

Первую назовем <u>главными опорными словами</u>. Они определяют главный предмет сообщения и по существу являются свернутым замыслом текста (Рилова, 1973, с. 35; Ваганова, 1974; Лингвистические вопросы..., 1983, с. 123). Такие слова и словосочетания в дальнейшем становятся своеобразным центр

ром, вокруг которого формируются другие текстовые элементы, отражающие составляющие микроситуации.

В каждой ситуации главные денотаты связаны с другими, менее важными элементами всей ситуации. Эти последние в то же время весьма важны для некоторых частей всей ситуации (микроситуаций) (Новиков, 1981, с. 53; Чистякова, 1979, с. 105). Соответствующие им слова текста назовем второстепенными опорными словами.

Формальными критериями выделения в тексте этих двух групп слов и словосочетаний являются: частота употребления лингвистической единицы во всем тексте (с учетом словарных и контекстуальных синонимов и местоименных замен) и количество абзацев, в которых встретилось слово или словосочетание (подробнее см.: Болдак, 1983).

В то же время, слова и словосочетания, входящие в каждую из двух выделенных групп, неоднородны по своему содержанию. В соответствии с предметными свойствами своих денотатов они образуют группы опорных слов-субъектов, объектов, слов-мест и слов-времени (Чистякова, 1979, с. 105; Лосева, 1973, с. 14-23). Именно они совместно с предикатами выражают взаимосвязи основных объективно существующих категорий: материи, движения, времени и пространства.

Поэтому уточняя определения понятия "предметно-логическое содержание абзаца" примем их следующие формулировки.

Назовем предметно-логическим содержанием абзаца-описания дифференцированное по типам главных и второстепенных субъектов, объектов, мест и времен ситуации, описанной в тексте, перечисление в абзаце фактов объективной действительности (признаков, явлений, событий, состояний, взглядов, теорий и т.п.), связанных с содержанием всего текста. Например, первый абзац военной корреспонденции К. Симонова "Полярной ночью" (Симонов, 1967, с. 357-361) записан так: "Это было в одну из долгих полярных ночей. Неровный серый свет сменивал все очертания; скалы, которые днем высоко громоздились над аэродромом, сейчас, казалось, мягко спускаются к нему, их очертания заволакивало мягкой дымкой". Это абзац-описание. Он описывает некоторую природную обстановку вокруг аэродрома. Слово "аэродром" является для всего текста главным опорным словом, описывающим место, где совершается действие текста. Предметно-логическое содержание этого абзаца-описания можно сформулировать так: "Описание природной обстановки вокруг главного места ситуации, пред-

ставленной в данном тексте".

Подробнее же образом определим понятия "предметно-логическое содержание" абзацев-повествований и абзацев-рассуждений.

Назовем предметно-логическим содержанием абзаца-повествования дифференцированное по типам главных и второстепенных субъектов, объектов, мест и времен ситуации, представленной в тексте, последовательное изложение в абзаце действий, событий, состояний с детализацией или без нее.

Аналогично предметно-логическим содержанием абзаца-рассуждения назовем дифференцированное по типам главных и второстепенных субъектов и объектов ситуации, представленной в тексте, последовательную связь в абзаце мыслей, связанных с содержанием всего текста.

Наконец, следует отметить еще одну особенность в организации содержания текста. Исследование структуры связных текстов показывает, что лексическое наполнение абзацев, их синтаксическое и грамматическое оформление, используемые средства связи в определенной степени зависят от места абзаца в общей структуре текста (Покусаенко, 1974; Змиевская, 1978, с. 13). Поэтому целесообразно при изучении содержания текстов выделять вводящие или начальные абзацы, внутренние или медиальные абзацы и абзацы заключительные или конечные.

В рамках построения алгоритма порождения текста нами были исследованы с рассмотренных выше точек зрения научные тексты по нейропсихологии и нейролингвистике А.Р. Лурия (62 текста), публицистические тексты К. Симонова (35 текстов) и поэтические тексты С. Есенина (172 текста)+. Общее соотношение числа абзацев с различным предметно-логическим содержанием в разных частях этих текстов представлено в таблице 1.

+ При анализе поэтических текстов строфа отождествлялась с абзацем прозаического текста (Акишина, 1979, с. 65; Сафарова, 1980; Гальперин, 1981, с. 55-57). Их исследование показало, что в них практически невозможно отделить начальные, медиальные и конечные строфы, так как все строфы таких текстов несут одинаковую смысловую нагрузку.

Таблица I

Соотношение числа типов абзацев
в различных частях трех видов текстов

Часть текста	Число различных типов абзацев					
	Тексты А.Р.Лурия		Тексты К.Симонова		Тексты С.Есенина	
	всего абзацев	разных	всего абзацев	разных	всего строф	разных
Начальная	78	32	47	29	-	-
Медиальная	960	52	1371	53	952	64
Конечная	60	24	30	16	-	-
Всего	1098		1448		952	

Рассмотрим детально типы абзацев всех трех видов текста по их предметно-логическому содержанию.

Наиболее употребительными среди начальных абзацев научного текста А.Р. Лурия оказались абзацы со следующим предметно-логическим содержанием*:

- "Утверждение о сути некоторого главного объекта" (М009, F = 9, f = 0,115)
- "Констатация некоторого состояния главного объекта в зависимости от других главных и второстепенных объектов" (М001, F = 8, f = 0,103)
- "Констатация некоторой специфичности главного объекта в связи с другими главными и второстепенными объектами" (М003, F = 8, f = 0,103)
- "Констатация некоторого прошлого действия автора" (М004, F = 5, f = 0,0640)
- "Констатация некоторого последующего действия автора, связанного с характеристикой главного объекта" (М024, F = 5, f = 0,0640).

По существу, это наиболее типичное содержание начальной части научного текста, представляющей главный объект исследования и то, что сделано или будет сделано автором для раскрытия сути этого объекта.

Подобный же анализ данных центральной части текстов

* В дальнейшем изложении для удобства описания вводятся: I) коды абзацев - Mijk, где ijk может принимать значения от 001 до 999; 2) абсолютная частота употребления типа абзаца - F; 3) относительная частота $f = \frac{F}{N}$, где N - общее число абзацев, описывающих каждую из трех частей соответствующего вида текста.

позволяет выделить следующие наиболее употребительные типы медиальных абзацев:

- "Утверждение об особенностях главного объекта с выходом на рисунок (таблицу и т.п.) и характеристика последнего" (МII6, F = II7, f = 0,I22)

- "Детальное описание поведения главного субъекта в связи с главным объектами (с примерами, выходами на рисунок, график и т.п. или без них)" (МII0, F = 83, f = 0,0864)

- "Подчеркивание характерной черты главного объекта и его влияния на главный субъект и другой главный объект" (МI09, F = 76, f = 0,0792)

- "Подчеркивание некоторой особенности главного объекта и краткое описание действий субъекта" (МII8, F = 63, f = 0,0656)

- "Констатация состояния главного объекта" (МI00, F = 42, f = 0,0438)

- "Констатация действий (состояния) главного субъекта" (МI05, F = 42, f = 0,0438)

- "Констатация прошлых и будущих действий автора" (МI07, F = 42, f = 0,0438).

Как видно, приведенное предметно-логическое содержание абзацев свидетельствует о том, что в средней, основной части текста, происходит сам процесс изучения главного объекта (МII6, МI09, МII8, МI00) путем анализа действий (поведения) главных субъектов (МII0, МI05) и некоторых действий (физических и умственных) автора, проводящего исследование МI07).

Наконец, среди абзацев, заключающих текст, наиболее типично следующее предметно-логическое содержание:

- "Заключительное утверждение о сути главного объекта" (М302, F = 9, f = 0,I50)

- "Констатация предстоящих действий автора" (М3I2, F = 9, f = 0,I50)

- "Утверждение о специфичности второстепенного объекта" (М3I8, F = 4, f = 0,067)

Такое содержание конечных абзацев вполне объяснимо: в заключении, как правило, формулируются основные выводы, результаты проведенного исследования (М302, М3I8) и намечаются действия автора по дальнейшему развитию исследования (М3I2).

Совместный анализ всех типов абзацев позволяет предположить, что в организации содержания научного текста, независимо от его структурной части, существуют определенные общие закономерности, выражающиеся в употреблении в разных частях такого текста абзацев с одним и тем же предметно-ло-

гическим содержанием. Степень общности начальной, медиальной и конечной частей научных текстов А.Р. Лурия по употребительности в них абзацев с одним и тем же предметно-логическим содержанием описывается данными, приведенными в таблице 2. Во всех трех частях этих научных текстов использовано 10 абзацев с одним и тем же предметно-логическим содержанием (соответственно 31,2 %, 19,2 % и 41,7 % от числа разных абзацев начальной, медиальной и конечной частей текстов).

Таблица 2

Употребляемость абзацев с одинаковым предметно-логическим содержанием в различных частях научных текстов А.Р. Лурия

Части текста	Части текста и количество одинаковых абзацев					
	Медиальная			Конечная		
	всего одинаковых абзацев	в % к числу начальных абзацев	в % к числу медиальных абзацев	всего одинаковых абзацев	в % к числу начальных абзацев	в % к числу медиальных абзацев
Начальная	21	65,6 %	40,4 %	13	40,6 %	54,2%
Медиальная	-			14	43,8 %	26,9%

Анализируя, по аналогии с научными текстами, особенности публицистических текстов К. Симонова по употребительности в них абзацев с различным предметно-логическим содержанием, можно отметить следующее.

Среди начальных абзацев этих текстов наиболее употребительны абзацы М4I6 ("Описание прочих объектов", F = 5, f = 0,I060), М4I7 ("Описание состояния автора и прочих субъектов", F = 4, f = 0,085I) и М4I8 ("Описание природы или внешней обстановки", F = 4, f = 0,085I). Они в определенной степени подтверждают высказывания исследователей о том, что начальная часть публицистического и художественного текста является некоторым прологом, раскрывающим предисторию того или иного события (Покусаенко, I974, с. I20; Змиевская, I978, с. I3). При этом происходит описание природы или внешней обстановки (М4I8), в которой будет проходить или проходило действие, окружающих предметов (объектов) (М4I6), представление каких-то сопутствующих основным персонажам действующих лиц (М4I6) и вполне естественное выражение отношения автора ко всему сказанному (М4I6).

Основой центральной части публицистических текстов являются I0 абзацев, накопленная частота которых составляет

62 % от общего числа всех абзацев, описывающих исследуемые публицистические тексты. По степени убывания частоты употребления к числу этих 10 абзацев относятся М508, М511, М519, М507, М545, М518, М506, М504, М523, М522 (см. табл. 3). И действительно, публицистические тексты описывают действия и состояния некоторых главных (М508, М511, М519, М507, М518, М523) и второстепенных персонажей (М522). В нем автор както характеризует свои главные персонажи (М506, М518, М523), предметы, являющиеся главными объектами описания (М504, М523), ту обстановку, в которой происходит развитие действий (М519, М545).

Среди абзацев, характерных для заключительной части публицистических текстов К. Симонова, можно выделить абзацы М700 ("Констатация утверждения о сути некоторого прочего объекта", $F = 5$, $f = 0{,}1670$), М706 ("Констатация будущих действий автора", $F = 3$, $f = 0{,}1000$), М709 ("Развернутое рассуждение автора, связанное с главными и второстепенными объектами и субъектами", $F = 3$, $f = 0{,}1000$) и М712 ("Высказывание автором предположения, связанного с главным субъектом", $F = 3$, $f = 0{,}1000$). Специфику заключительной части таких текстов можно определить как некоторое послесловие или итог вышесказанному, проявляющийся в констатации сути некоторых объектов (М700), в высказывании заключительных рассуждений автора о том, что было описано в центральной части текста (М709) или предположения о будущих судьбах главных персонажей (М712). Наконец, здесь же формируются последующие действия автора, связанные с развитием темы или её завершением (М706).

Если сравнить между собой все абзацы, использованные при создании публицистических текстов К. Симонова, то можно отметить, что в отличие от научных текстов, здесь существует определенная специфика в построении начальной, центральной и заключительной части таких текстов. Всего нами зафиксировано лишь 4 типа абзацев, которые оказались общими для всех 3-х частей текстов. Это составляет соответственно 13,8 %, 7,5 % и 25 % от числа всех абзацев, использованных при создании начальной, центральной и конечной частей текстов.

Одиннадцать типов абзацев оказались общими для начальной и центральной частей текста, 6 - для центральной - конечной и 8 - для начальной - конечной частей текста (см. табл. 4).

5

Таблица 3

Наиболее употребительные типы абзацев медиальной части текстов К. Симонова

№ п/п	Код	Тип абзаца	Частота F	Частота f
I.	М504	Описание главного объекта и прочих субъектов	42	0,0306
2.	М506	Краткая характеристика главного субъекта	48	0,0350
3.	М507	Констатация последовательных действий двух главных субъектов	63	0,0459
4.	М508	Констатация некоторых действий главного субъекта	218	0,159
5.	М5II	Описание состояния главного субъекта и его действий	I67	0,I22
6.	М5I8	Характеристика главного субъекта и констатация его последующих действий	57	0,04I6
7.	М5I9	Описание внешней обстановки и констатация действий главного субъекта	I22	0,0890
8.	М522	Описание действий второстепенного субъекта	35	0,0255
9.	М523	Констатация действий главного субъекта и характеристика главного объекта	40	0,0292
I0.	М545	Констатация некоторой внешней обстановки	59	0,0430

Таблица 4

Употребляемость абзацев с одинаковым предметно-логическим содержанием в различных частях публицистических текстов К. Симонова

части текста	Части текста и количество однотипных абзацев				
	Центральная			Конечная	
	Всего	в % к числу начальных	в % к числу медиальных	всего	в % к числу начальных	в % к числу медиальных или конечных
Начальная	II	37,9 %	20,8 %	8	27,6 %	I5,I %
Центральная	-			6	20,7 %	II,3 %

Как видно из таблицы 4, существует определенная специфика в организации различных частей публицистических текстов. С этой точки зрения научный текст можно считать более однородным (ср. табл. 2, стр.32).

Анализируя типы строф поэтических текстов, можно выде-

лить следующие 7 строф, которые являются наиболее употребительными для поэтических текстов С. Есенина: М8I7, М806, М85I, М802, М805, М8I8 и М838 (см. табл. 5). Их накопленная частота составляет 49,7 % от числа всех 952 строф, использованных для исследования.

В совокупности предметно-логическое содержание этих строф свидетельствует о том, что в текстах С. Есенина описываются состояния или некоторые действия автора (М8I7, М806, М805, М85I, М802, М8I8, М838), состояния или действия главных персонажей (М8I8, М85I), состояние природы (М806).

Как видно из сказанного, эти тексты значительно отличаются по своему содержанию от публицистических текстов, где ведущую роль в их медиальной части играют действия и состояния главных и второстепенных персонажей, а также определенные объекты описания и та обстановка, в которой происходят указанные действия (ср. стр. 33).

Еще более значительно отличие содержания текстов С.Есенина от научных текстов А.Р. Лурия. В последних, как мы отмечали выше (стр.30 - 3I), речь идет об особенностях некоторых исследуемых объектов и поведении некоторых субъектов, определяющим эти особенности объектов.

Таблица 5

Наиболее употребительные типы строф поэтических текстов С. Есенина

№ № п/п	Код	Тип строфы	Частота F	Частота f
I.	М802	Констатация состояния или действия автора, связанных с главным субъектом	57	0,0600
2.	М805	Описание состояния автора	44	0,0462
3.	М806	Описание природы и констатация действия или состояния автора	9I	0,0956
4.	М8I7	Описание состояний и действий автора	I46	0,I534
5.	М8I8	Констатация действия главного субъекта и будущего действия автора	40	0,0420
6.	М838	Констатация вопросов автора и описания состояния автора	34	0,0357
7.	М85I	Побуждение автора к выполнению действий главным субъектом и констатация некоторых действий и состояний автора	6I	0,064I

Таким образом, знание статистических особенностей употребления абзацев с различным предметно-логическим содержанием позволяет четко отличить тексты различных функциональ-

ных стилей. Выделенные типы абзацев являются также основой для изучения содержания конкретных текстов как некоторой последовательности таких абзацев. Однако для этого необходимы данные о взаимном расположении выделенных разновидностей абзацев. Выяснение этого - задача отдельного исследования.

Л И Т Е Р А Т У Р А

Адмони В.Г. Синтаксическая семантика - это семантика синтаксических структур. - В кн.: Проблемы синтаксической семантики. Материалы научной конференций. М., 1976, с.3-8.

Акишина А.А. Структура целого текста. Выпуск II. - М., 1979.

Анохин П.К. Опережающее отражение действительности. - Вопросы философии, 1962, № 7.

Бахтин М.М. Проблема речевых жанров. - В кн.: Бахтин М.М. Эстетика словесного творчества. М., 1979, с. 237-280.

Болдак И.А. Специфика развития темы в научном тексте (в печати).

Бондаренко Г.В., Шрейдер Ю.А. Текст - смысл - ситуация (к постановке проблемы). - В кн.: Вопросы информационной теории и практики. № 36. М., ВИНИТИ, 1978, с. 80-91.

Вахенина В.П. К вопросу о лингвистических предпосылках смыслового анализа текстов на языковых факультетах. - В кн.: Проблемы лексической и грамматической семасиологии. Владимир, 1974, с. 194-205.

Васильев Ю.А. О влиянии композиционно-смысловой организации научного текста на его языково-стилистические характеристики. - В кн.: Стиль научной литературы. М., 1978.

Верещагин Е.М., Костомаров В.Г. Лингвострановедческая теория слова. - М., 1980.

Гальперин И.Р. Интеграция и завершенность текста. - Изв. АН СССР. Сер. лит. и языка, 1980, т. 39, № 6, с. 512-520.

Глаголев Н.В. Семантико-структурные элементы актуализации сообщения в сверхфразовом единстве. - В кн.: Проблемы синтаксической семантики. Материалы научной конференции. М., 1976, с. 104-107.

Гришина О.Н. Проблемы контекстно-вариативного членения текста в стиле языка художественной и научной прозы. - В кн.: Функциональные стили и преподавание иностранных языков. - М., 1982, с. 52-68.

Денисов П.Н. Принципы отбора лексики для учебных словарей. - В кн.: Вопросы учебной лексикографии. М., 1969.

Егоров В.Л. Описание конструкции технического объекта как один из подвидов речевой формы "описание". - В кн.: Функциональные стили и преподавание иностранных языков. М., 1982, с. 42-52.

Зарубина Н.Д. К вопросу о природе сложного синтаксического целого. - В кн.: Вопросы психолингвистики и преподавание русского языка как иностранного. М., 1971.

Зарубина Н.Д. О психолингвистическом обосновании приемлимости предложения и сверхфразового единства в качестве единиц обучения иностранному языку. - В кн.: Психология грамматики. М., 1968, с. 145-152.

Змиевская Н.А. Лингвостилистические особенности дистантного повтора и его роль в организации текста (на материале английской и американской прозы). - Автореф. дис. ... канд. филол. наук. М., 1978.

Кацнельсон С.Д. Типология языка и речевое мышление. - Л., 1972.

Кестен С.А. Структура абзаца и его роль в архитектонике и композиции художественного текста. – Иностранные языки в вузах Узбекистана. Вып. 8. Ташкент, 1975, с. 256-264.
Кожин А.Н., Крылова О.А., Одинцов В.В. Функциональные типы русской речи. – М., 1982.
Колшанский Г.В. Категория семантики в синтаксисе. – В кн.: Проблемы синтаксической семантики. Материалы научной конференции. М., 1976, с. 246-248.
Колшанский Г.В. О языковом механизме порождения текста. – ВЯ, 1983, № 3, с. 44-51.
Ларина А.А. О предметно-логическом содержании абзацев (на материале английских текстов по автотранспорту). – В кн.: Проблемы внутренней динамики речевых норм. Сборник научных статей. Минск, 1982, с. 203-209.
Лингвистические вопросы алгоритмической обработки сообщений. – М., 1983.
Ломоносов М.В. Полное собрание сочинений. Том 7. – М.,1953.
Лосева Л.М. Структурно-семантическая организация целых текстов (методические рекомендации учителям-словесникам по изучению связной речи). – Одеса, 1973.
Мильчин А.Э. Методика редактирования текста. – М., 1980.
Нечаева О.А. Функционально-смысловые типы речи (описание, повествование, рассуждение). – Улан-Удэ, 1974.
Новиков А.И., Чистякова Г.Д. К вопросу о теме и денотате текста. – Изв. АН СССР. Серия лит. и языка, 1981, том 40, № 1, с. 48-56.
Перекальская Т.К. Абзац как основная единица научного текста. – В кн.: Вопросы методики и лингвистики. М., 1980, с. 8-10.
Пешковский А.М. Русский синтаксис в научном освещении. – М., 1920.
Покусаенко В.К. К вопросу о функциях абзаца. – В кн.:Структура предложения и абзац. Ростов-на-Дону, 1974, с. 112-121.
Рылова Т.Н. Принципы упорядочения и группировки значимых слов текста для совершенствования критерия семантической связи между предложениями. – Научно-техническая информация, сер. 2, № 4, 1973, с. 34-38.
Сафарова А.С. Семантическая композиция текстов английских лирических стихотворений. – В кн.: Сборник научных трудов МГПИИЯ. Вып. 155. М., 1980, с. 75-86.
Сахарный Л.В. Типология структуры текста с точки зрения теории речевой деятельности. – В кн.: Лингвистика текста. Материалы научной конференции. Часть II. М., 1974, с. 39-45.
Серкова Н.И. Семистика абзаца в научной и художественной литературе. – В кн.: Семантические вопросы микро- и макросинтаксиса. Хабаровск, 1980, с. 109-117.
Симонов К. Полярной ночью. – В кн.: Собрание сочинений в шести томах. Том 2. М., 1967, с. 357-361.
Смирнов А.А. Проблемы психологии памяти. – М., 1966.
Солнцев В.М. Язык как системно-структурное образование. – М., 1977.
Страхова З.С. Внешние средства организации текста. – В кн.: Лингвистика текста. Сборник научных трудов МГПИИЯ. Вып. 141, М., 1971, с. 150-159.
Узнадзе Д.Н. Экспериментальные основы психологии установки. – В кн.: Психологические исследования. М., 1966.
Чесноков П.В. Об отношении между речевым и мыслительным процессами с точки зрения единства языка и мышления. – В кн.: Синтаксическая семантика и прагматика. Калинин, 1982, с. 38-47.

Чистякова Г.Д. Смысловая структура текста как определяющий фактор его понимания. - В кн.: Семантика, логика и интуиция в мыслительной деятельности человека. М., 1979, с. 101-126.

Koen F., Becker A., Young R. Psychological reality of the paragraph. - In: Studies in language and language behavior. The University of Michigan, VI, 1968.

Stalnaker R.C. Pragmatics. - In: Semantics of natural language. Dordrechts, 1972, pp. 380-397.

SPECIFIC FEATURES OF RUSSIAN TEXTS VIEWED AS PARAGRAPH USAGE WITH DIFFERENT SUBJECT-LOGICAL CONTENTS

Alexandr V. Zubov

S u m m a r y

A possibility of paragraph usage of a prose text and strophe usage of a poetic one as the main semantic-syntactic text unit is grounded in the article. A notion of "subject-logical contents of a paragraph (strophe)" is introduced on the basis of the main and secondary text supporting units. Paragraph (strophe) types are revealed according to their subject-logical contents on the basis of scientific texts by A.R. Luria, publicistic texts by K. Simonov and poems by S. Esenin. Peculiarities of the usage of these paragraph (strophe) types are presented in the opening, intermediate and final parts of the texts under analysis.

БАНКИ ДАННЫХ 80-Х: ТЕОРИЯ, ЭКСПЕРИМЕНТ, ВНЕДРЕНИЕ

Р.Ю. Кобрин

Джеймс Мартин, автор широко известной в мире монографии "Организация баз данных в вычислительных системах" (Мартин, 1980), назвал 70-е годы нашего столетия десятилетием баз данных: "Разработка баз данных общего пользования останется на долгие годы одним из основных направлений деятельности в области обработки данных ... Во всех областях жизни и производства банки данных изменят характер деятельности человека. Историки будут рассматривать появление банков данных на ЭВМ и возможностей, связанных с ними, как шаг, изменивший природу эволюции общества и имеющий, возможно, большее значение, чем изобретение печатного станка" (Мартин Дж., 1980).

Проектирование банков данных является новым этапом в развитии автоматизированных систем обработки информации. Этот этап характеризуется большими объёмами и сложностью структур обрабатываемой информации, сокращением времени и стоимости информационного обслуживания, разработкой специальных программных комплексов управления данными на ЭВМ.

Под автоматизированным банком данных (БД) обычно понимается организационно-техническая система, представляющая собой совокупность баз данных пользователей, технических и программных средств формирования и ведения этих баз и коллектива специалистов, обеспечивающих функционирование системы.

Центральная проблема создания банка данных: конструирование базы данных, представляющей собой поименованную совокупность данных, отображающую структуру объекта и отношения между объектами (Шёрс, 1978).

Важно подчеркнуть, что банк данных - прежде всего, информационная система и необходимо определить место БД среди других информационных систем, а также определить структуру БД, методы построения БД, наметить перспективные пути развития БД как информационных систем. Иначе говоря, должна быть построена теория БД.

Под "теорией" обычно понимается комплекс идей, направленных на истолкование и объяснение какого-либо явления. Теория должна дать целостное представление о закономерностях и существенных связях определённой области действительности - объекта теории. Центральную роль в формировании теории играет лежащий в её основе идеализированный объект - теоретичес-

кая модель. Эта теоретическая модель может предполагать или не предполагать математического описания.

Всякая теория начинается с конкретизации объекта исследования. В.И. Ленин подчёркивал, что теория должна дать "объект в его необходимости, в его всесторонних отношениях". Таким образом, одна из главных задач теории - определение объекта исследования.

Если с этих позиций мы подойдём к существующей практике построения БД, то вопрос необходимо сформулировать так: "Что является объектом исследования в теории банков и баз данных? Ответ на этот вопрос отнюдь не прост. Сказать, что объектом теории БД являются существующие банки и базы данных, было бы не совсем правильным.

Конечно, опираясь на теорию, человек способен создавать то, что не существует в природе или социальной действительности, но, строго говоря, эти объекты являются результатами теоретической и практической деятельности.

Необходимо выделить две стороны теоретической деятельности. С одной стороны, анализируя созданные объекты, возможно создавать теории, объясняющие объекты и прогнозирующие их развитие. С этой точки зрения можно сказать, что объектами теории БД являются существующие БД. Но, с другой стороны, объектом теории БД должны являться информационные процессы, прежде всего, данные, участвующие в этих процессах, а также - и это особенно важно - стандартные информационные ситуации действительности. В последнем случае говорят не о банках данных, а о банках (базах) знаний. Информационные процессы и ситуации являются объектом теории БД постольку, поскольку они должны структурироваться и формализоваться для представления в БД. Информационные процессы и ситуации являются объектом и других наук - например, теории информации, социальной психологии, социологии, но при этом выделяются различные аспекты рассмотрения.

А теперь к БД. Есть ли у нас основания утверждать, что наука располагает теорией БД?

В последние годы появилось несколько крупных зарубежных и отечественных работ, обобщающих различные концепции систем банков и баз данных. (Кузин Л.Т., 1976; Клыков Ю.И., Горьков Л.Н., 1980; Овчаров Л.А., Селетков С.Н., 1982; Кокорева Л.К., Малашинин И.И., 1984; Дейт К., 1980; Мартин Дж., 1980; Шенк Р., 1980; Олле Т., 1981; Атре Ш., 1983; Wells, Hopkinson, 1983; Codd, 1984).

Так, в работе (Wells, Hopkinson, 1983) формулируется общая идеология организации информации по принципу БД, вводятся основные понятия БД, описывается архитектура БД, формулируется понятие модели данных.

Однако и в этой и в других работах не установлены отношения между БД и другими информационными системами, не конкретизировано понимание концептуальной модели данных, не установлены связи между концептуальной моделью и лингвистическим представлением данных. Можно сказать, что существует обширная практика построения систем данных, но отсутствует теория, объясняющая объекты науки, показывающая их место среди других фактов действительности, прогнозирующая их развитие.

Адекватное описание информационных процессов требует построения математических моделей, характеризующих эти процессы. Сложившаяся в области математического моделирования ситуация напоминает современное развитие физики: когда выяснилось, что линейная физика не может объяснить некоторые природные явления, возникла теория, получившая название нелинейной физики. Известно, что информационные процессы, в отличие от энергетических, дискретны, и для их описания целесообразны дискретные математические модели. Однако сложные информационные процессы, в которых чуть ли не решающую роль играет психология социального коллектива, с трудом "укладываются" в рамки дискретных математических моделей.

Таким образом, с этих позиций трудно говорить о теории банков данных. Но банки данных развиваются, хотя и не столь быстрыми темпами, как принято считать. Авторы (Heidlerová K., Chvalovský V., 1983) утверждают, что в настоящее время в США только 10-20 % всей информации, необходимой для управления производством, обрабатывается с использованием стандартных систем управления базами данных (СУБД) и пакетов прикладных программ, т.е. методами БД. В других странах, указывают авторы, этот процент ещё ниже, хотя прогнозы, сделанные 5 лет назад, указывали на более значительные цифры.+

Развитие банков данных затруднено не только, а может быть и не столько тем, что переход на СУБД требует значительных затрат человеческих и материальных ресурсов. Дело в

+ По данным (Datenbankführer, 1983) в ФРГ действует 86 крупных коммерческих банков данных, что, конечно, не полностью удовлетворяет потребности производства.

том, что "теория проектирования банков данных далека от своего завершения и находится в процессе становления" (Кокорева Л.В., Малашинин И.И., 1984).

Внимательное рассмотрение теории и практики БД позволяет увидеть интересный парадокс, на который впервые обратили внимание в работе (Кокорева Л.В., Малашинич И.И., 1984).

Сейчас, как говорилось выше, в теории БД отсутствуют необходимые формализмы. Для описания данных произвольной природы и манипуляций с ними пытаются использовать различные алгебраические методы, прежде всего методы аппликативных вычислительных систем, широко применяются реляционные модели. Информационная потребность (ИП) пользователя выражается в языке запросов. Сама по себе ИП - категория психологическая и семантическая, не поддающаяся формализации.+ Попытка точной формулировки ИП в языке запросов приводит к тому или иному варианту исчисления предикатов. Но эксплуатация БД очень скоро показывает, что реальные фрагменты действительности несводимы к их формальным моделям. Поэтому "любая формализация проектирования банка данных явно или неявно включает в себя определённые содержательные описания моделируемых фрагментов реальной исследуемой предметной области." (Кокорева Л.В., Малашинин И.И., 1984).

Предметом нашего рассмотрения являются информационные системы, в которых функционирует т.н. семантическая информация, т.е. информация, выражающаяся средствами ЕЯ. Хотя в принципе все сказанное может быть отнесено и к системам иного типа, например, системам автоматизированного проектирования (САПР), в которых обрабатывается информация, выражаемая сложным конгломератом знаковых средств, объединяющих ЕЯ, искусственные семиотические системы (например, язык топографических и иных видов карт, чертежей, проектов и т.д.), математические символы (Васин Ю.Г. и др., 1983).

Примерно 10-15 лет назад довольно четко выделялись типы информационных систем: автоматизированные системы управления, информационно-поисковые системы различного вида (Кобрин Р.Ю., 1972; Кобрин Р.Ю., Максимов В.Р., 1975; Аграев В.А. и др., 1975). Не вполне определённой была структура АСУ и, в

+ Даже при формулировании собственной ИП в терминах естественного языка (ЕЯ) утрачивается часть информации. Можно вспомнить Тютчевское "Мысль изречённая есть ложь".

частности, место ИПС в АСУ. В середине 70-х годов было установлено, что различные подсистемы АСУ могут располагать собственными информационно-поисковыми системами, хотя в принципе возможна и разработка единой ИПС, обслуживающей АСУ в целом (Аграев В.А. и др., 1975).

Сомнений не вызвало и следующее обстоятельство: обеспечивающими подсистемами АСУ являются информационное, математическое и техническое обеспечение. Казалось - по крайней мере, с позиций теоретических - что задача разработки информационного обеспечения - это прерогатива специалистов по прикладной лингвистике. Однако, в действительности информационным обеспечением занимался кто угодно: системщики, математики, специалисты-отраслевики, экономисты, но не лингвисты. Многолетняя практика разработки АСУ различных типов показала, что - несмотря на достаточно продвинутую теорию АСУ - конкретных жизнеспособных АСУ, особенно АСУ отраслей, крупных предприятий, призванных обрабатывать сложноструктурированные семантически насыщенные данные, так и не было создано.

Внимательный анализ литературы показывает, что сейчас сам термин АСУ редко употребляется в литературе. Совершенно очевидна тенденция к его вытеснению терминами "банк данных", "база данных", "информационная система", "система обработки данных". В западных странах термин АСУ не используется, что, конечно же, не означает отсутствия самих систем. Любая коммерческая организация осуществляет учет и контроль, невозможный без использования средств ВТ. Например, широко известная система кредитных карточек в США, заменяющая денежные ассигнации, основывается на учете, осуществляемом кредитными и прочими банковскими учреждениями США.

Концепция банков и баз данных начала складываться в конце 60-х годов (Engles, 1972). Как пишет Дж. Мартин (Мартин Дж., 1980), "как часто бывает, когда новое понятие становится модным, многие пользователи начали применять его к своим файлам, изменив только название, но не изменяя при этом их свойств". Эволюция систем обработки данных была связана в первую очередь с появлением ЭВМ 3,4,5 поколений и совершенствованием средств программного обеспечения. Из американской литературы к нам проникли и стали широко использоваться, часто применительно к морально устаревшим системам, термины "банк данных", "база данных", "система управления базой данных".

И вот здесь возникла чрезвычайная путаница с терминологией. В отечественной литературе: "банк", "база", ИПС,ИСС, АСУ, САПР, АСПР, АСУ ТП, АСОД и т.д. В каких отношениях эти понятия находятся друг к другу, должны ли локальные ИПС образовывать банк, или же банк рассматривается как одна ИПС - не ясно. Путаница с терминологией существует и на Западе. В работе (Мартин Дж., 1980) отмечено, что некоторые специалисты вместо термина "база данных" используют термин "банк данных", а под базой данных подразумевают совокупность банков данных. К. Дейт также отмечает (Дейт К., 1980), что, как и многие другие новые области, область систем баз данных не обладает еще общепринятой терминологией. Известно, что ассоциация CODASYL, объединяющая разработчиков БД, большое внимание уделяет нормализации терминологии: проводится согласование терминологий различных фирм, издаются словари, разрабатываются терминологические ГОСТы.

Неудовлетворительность терминологической ситуации удобно проиллюстрировать на примере информационного обеспечения (ИО) - важнейшей составной части БД.

ГОСТ 19675-74 определил ИО как совокупность единой системы классификации и кодирования технико-экономической информации, унифицированных систем документации и массивов информации, используемых в АСУ. Этот же ГОСТ определяет лингвистическое обеспечение (ЛО) АСУ как "совокупность научно-технических терминов и других языковых средств, используемых в АСУ, а также правил формализации ЕЯ, включая методы сжатия и развёртывания текстов в целях повышения эффективности машинной обработки информации и облегчающие общение человека с машиной".

Из этих определений трудно понять, каковы действительные отношения между ИО и ЛО? Между тем в типичной работе, описывающей широко известную СУБД СЕДАН (Валькман Ю.Р., 1980) сказано, что "подсистеме информационного обеспечения любой системы обработки данных необходимо решить две проблемы: во-первых, выбрать СУБД, во-вторых, спроектировать программное обеспечение, позволяющее использовать её при создании конкретной системы".

Даже в сборнике "Итоги науки и техники. Информатика" (-М.: ВИНИТИ, 1981) программные средства манипулирования с данными и информационными массивами, позволяющие пользователю непосредственно работать с системой в интерактивном режиме, включены в лингвистическое обеспечение БД.

Вся эта путаница еще одно свидетельство неразработанности теории БД и крайнего невнимания к вопросам собственно информационно-лингвистического обеспечения, непонимания его специфики.

В.М. Глушков и Ю. Каныгин писали в газете "Правда": "На первое место выдвигаются информационные, организационно-экономические, социальные проблемы. Известный "техницизм" в формировании АСУ, закономерно доминировавший на первых этапах "электронизации" пора изжить. Дело ведь не в том, чтобы повсюду настроить ВЦ и соединить их каналами связи ... Такую сеть, лишенную соответствующей информационной "начинки", с точки зрения хозяйственной полезности можно было бы уподобить египетским пирамидам" ("Правда", I3.XII.I98I г.).

Под базой данных понимается "упорядоченная совокупность однотипных данных, представленных обычно в машиночитаемой форме ... и относящихся к какой-либо отрасли, теме или предмету" (РЖ "Информатика", I98I, № I). Существует полезное понятие "архитектура баз данных", имеющая внешний, внутренний и концептуальный уровни. В отечественной литературе этим терминам соответствуют термины "логический", "физический", "семантический" уровни. Подавляющее большинство существующих систем ("Ока", СИНБАД, СИОД-I, СИОД-2, "Кама") характеризуются логическим и физическим уровнями. Причем логический уровень рассматривается безотносительно к семантическому содержанию данных. К. Дейт отмечал, что "очень немногие современные системы действительно поддерживают концептуальный уровень" (К. Дейт, I980). В.И. Будзко также подчёркивает, что "возможность задания концептуальной модели в большинстве современных СУБД отсутствует" (Будзко В.И., I983).

Концептуальному уровню соответствует концептуальная модель, представляющая полное информационное содержание базы данных, которая: I) состоит из перечня управляемых объектов и отношений между ними, 2) основывается на логико-математическом структурировании предметной области, 3) м.б. представлена в виде "концептуальной записи" (графа, тезауруса, фрейма).

Результатом логико-математического структурирования предметной области должна стать математическая модель, призванная - на основе познания сущности объекта - формально описать его структуру. Ситуация осложняется тем, что: I) сущность многих моделируемых объектов не познана, 2) име-

ются принципиально (для сегодняшнего уровня развития науки) неформализуемые объекты. Поэтому появляются математические модели, неадекватно описывающие объект, что приводит к невозможности практического использования системы.

Удобной формой представления концептуальной модели является граф. Граф позволяет наглядно представить и формализовать понятия предметной области и отношения между ними, фиксированные в рамках системы лингвистического обеспечения БД в классификаторе информации. Лингвистической основой классификатора является тезаурус-словарь, в котором адекватно (с учётом классификационных связей) представлена система понятий определённой предметной области.

Построение любого тезауруса основывается на следующих элементарных предпосылках:

I. В непрерывном континууме объективной действительности существуют предметы (объекты), признаки и отношения, объединённые в ситуации.

2. Язык дискретизирует ситуации и их элементы.

3. Наука - в процессе познания действительности и при помощи языка - также дискретизирует ситуации и элементы.

Можно сказать, что в объективной действительности существуют предметы, признаки и отношения, образующие ситуации, дискретизируемые познавательной деятельностью человека и обязательно отражаемые в языке (естественном или искусственных). Всякий объект: I) системен, 2) назван в языке, 3) является информационным объектом. Создание концептуальной модели состоит в выявлении и экспликации языковых единиц, называющих составные части объекта и отношений между ними.

Процесс создания концептуальной модели удобно иллюстрировать на примере спортивных игр. Пусть управляемым объектом будет хоккейная команда. Для человека, не знакомого с хоккеем, она представляет собой нерасчленённый класс молодых людей, занимающихся непонятными действиями с неясными орудиями. Познание сущности объекта ведёт к его дискретизации и названию составных частей. Есть термины: вратарь, защитник, нападающий, тренер, тройка, вбрасывание, шайба, айсинг и т.д.; существуют отношения между элементами объекта, также имеющие своё языковое выражение. Для построения концептуальной модели необходимо разработать перечень лексических единиц и конкретизировать отношения. Модель должна быть динамичной, т.е. предусматривать описание цели дейст-

вия, всех возможных ситуаций. Модель может быть либо тезаурусной: ориентироваться на семантическое представление предметной области в целом, либо фреймовой: ориентироваться на адекватное представление конкретных актуальных информационных ситуаций.

Современная терминология и прикладной характер задач не должны "вуалировать" очевидный теоретико-лингвистический характер концептуальных моделей. Концептуальные модели - это, по сути дела, формализованные семантические поля, теория которых разрабатывалась в прошлом крупнейшими лингвистами - Гумбольдтом, Л.В. Щербой, зарубежными и отечественными языковедами - Вайсгербером, Трио́м, Найдой, Ю. Апресяном, Ю. Карауловым, В. Морковкиным, Э. Скороходько. Тем удивительнее тот факт, что в нашей стране как пионерская воспринимается теория фреймов М. Минского. Но, строго говоря, фрейм - это языковое описание ситуации, т.е. тезаурус ситуации, и принципиальных различий во фреймовом и тезаурусном моделировании нельзя усмотреть.

Построение тезаурусной концептуальной модели - задача чрезвычайно сложная. Большинство существующих тезаурусов - перечни (и далеко не всегда полные) объектов с чрезвычайно бедным набором отношений (Кобрин Р.Ю., 1979).

С чем это связано?

1. Очень трудно выделить термины, описывающие объекты и ещё труднее ранжировать их по степени значимости. Известно, что в сознании специалиста любой отрасли знания отражена система соответствующих понятий. Но если в процессе психолингвистического эксперимента попросить испытуемых назвать базовые термины предметной области, дать им определения и установить отношения между ними, то эта задача вызовет затруднения. Это экспериментально доказанный факт. Дело в том, что между фактами человеческого мышления (понятиями) и словами-терминами существует определённый разрыв, приводящий к тому, что в процессе коммуникации мы утрачиваем часть хранящейся в мозгу информации.

2. Современная логика не имеет чёткой и непротиворечивой теории отношений.

Разработка тезауруса - это всегда работа с языком и с предметной областью, это логико-лингвистическое моделирование предметной области (Кобрин Р.Ю., 1979).

Не претендуя на решение сложных теоретических проблем БД, попытаемся наметить основные этапы создания автоматизи-

рованного банка данных:

I. Анализ информационных данных.

Логико-лингвистическое моделирование предметной области с использованием методов прикладной лингвистики и логики.

2. Построение концептуальной модели предметной области:

- построение тезауруса,
- построение классификатора информации,
- представление тезауруса в виде реляционной модели.

3. Выбор или построение системы управления базой данных (СУБД).

4. Отображение концептуальной модели в логической модели, обеспечиваемой в СУБД.

5. Проектирование физической модели и её оценка.

6. Внедрение и обеспечение требуемых эксплуатационных характеристик (Атре Ш., 1983).

Намеченные этапы соответствуют в основном главным компонентам банка данных. Принято говорить, что БД - это база данных + система управления базой данных.[+] Иными словами, БД - это концептуальная модель --> логическая модель --> физическая модель, реализованные на ЭВМ современными программно-техническими средствами.

С точки зрения обеспечивающих подсистем можно выделить:

I. Информационно-терминологическое обеспечение (ИТО), включающее в себя лингвистические средства для описания данных (различного рода тезаурусы, классификаторы, рубрикаторы и словари), а также необходимые средства предмашинной подготовки информации.

ИТО может рассматриваться как информационно-поисковый язык, обеспечивающий однозначное представление данных и их последующее хранение и поиск для решения информационных задач пользователей.

В рамках ИТО создаётся концептуальная модель базы данных.

2. Программное обеспечение (ПО), предназначенное для физического ввода/вывода данных и их обработки, автоматизи-

---

[+] При таком понимании нивелируется роль человека-"эксплуататора" и пользователя, а БД - система прежде всего человеко-машинная.

рованного ведения словарей и классификаторов, обновления состава данных в БД, моделирования, планирования и прогнозирования при решении информационно-логических задач.

Программное обеспечение реализовано в виде пакетов прикладных программ и является основной частью СУБД.

3. Технико-технологическое обеспечение включает в себя комплексы ЭВМ, терминалов, средств связи, различных устройств ввода-вывода и обеспечивает комплектование БД информацией, первичную аналитическую обработку и индексирование информации, ввод данных в ЭВМ, функционирование ЭВМ, отображение и размножение выходной информации, связь и телекоммуникацию внешних пользователей.

В традиционной терминологии информационного поиска этим понятиям соответствуют термины "лингвистическое", "информационное", "математическое" и "техническое" обеспечение.

Следует ещё раз подчеркнуть, что центральная роль в БД принадлежит концептуальной модели, которая служит средством общения между различными пользователями и не зависит (и не должна зависеть) от применяемой СУБД. Проектирование логической и физической модели, напротив, в значительной степени определяется СУБД (Атре Ш., 1983).

Часто приходится сталкиваться с пожеланиями пользователей: "Поставьте нам банк данных "Ока", "Кама", ИНЭС. НАБОБ ...". Но это не банки данных, а СУБД, и в каждом случае необходимо построение концептуальной модели данных и согласование её с СУБД. На практике чаще всего СУБД определяется заранее, но лучше всего осуществлять выбор СУБД после того, как спроектирована концептуальная модель (Атре Ш., 1983).

Рассмотрим основные процедуры построения БД на примере банка картографической информации, созданного в Горьковском НИИ прикладной математики и кибернетики.

Первый этап создания БКД (банка картографических данных) состоит в разработке концептуальной модели, основывающейся на представлении картографической информации набором единиц трёх типов:

а) словами и словосочетаниями ЕЯ,

б) системой условных знаков,

в) математическими символами (Васин Ю.Г. и др., 1983)

Технологически концептуальная модель реализована в информационно-терминологическом обеспечении (ИТО), представляющем собой интегральный аппарат информационного и лингв-

вистического обеспечений, который позволяет:

-реализовать описание семантики условных знаков (информационное обеспечение),

-реализовать семантику терминологической системы карты (терминологическое обеспечение).

ИТО БКД имеет следующую структуру:

I. Базовая структура (БС).

2. Система предмашинной подготовки (СПП).

3. Классификатор единиц информации (КЕИ).

4. Языковое обеспечение диалога (ЯОД).

Структурная схема функционирования ИТО представлена на рис. № I

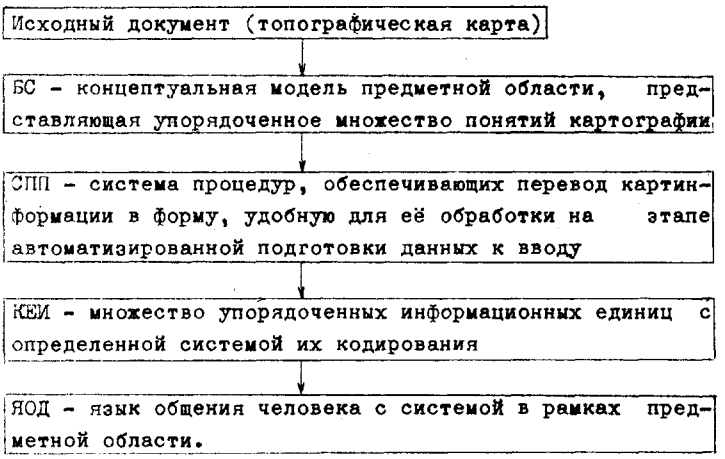

Рис. № I Схема функционирования ИТО БКД

Второй этап состоит в отображении ИТО (концептуальной модели) в логической модели. В БКД этот процесс реализован в виде процедурных грамматик, позволяющих структурировать объекты на логическом уровне (Бородин В.В., I983).

На третьем этапе происходит переход к физической модели в рамках СУБД. Этот переход может происходить в рамках СУБД ИНЭС (Ясаков Ю.В., I983) или вновь созданной СУБД (Васин Ю.Г., Кустов Е.А. и др., I983).

Четвертый этап призван отработать оптимальный режим диалога и обеспечить требуемые эксплуатационные характеристики.

Разработанный БКД прошёл опытную эксплуатацию и используется в настоящее время в рамках автоматизированной

системы научных исследований "Акварель" (Васин Ю.Г., Кустов Е.А. и др., 1983).

Причина трудностей и неудач в построении и эксплуатации банков данных, как представляется, лежит в практическом отсутствии концептуальных моделей в большинстве БД. Строго говоря, многие современные БД - это системы управления базами данных, причём вопросам соответствия информационных форматов СУБД реальной структуре данных, и что особенно важно, информационной потребности пользователя не уделяется должного внимания. Остаются нерешёнными и многие социопсихологические проблемы взаимодействия пользователя и системы. Очевидно, разработкой концептуальных моделей и информационно-терминологического обеспечения БД в целом должны заниматься специалисты данной предметной области, осуществляющие предварительное структурирование предметной области, лингвисты-прикладники, занимающиеся логико-лингвистическим моделированием, и математики - обеспечивающие переход от логико-лингвистической модели к математической модели предметной области, если такой переход возможен.

Проблема построения теории БД, разработки теоретико-лингвистического обоснования концептуальных моделей БД, конкретных принципов и методик их построения ждёт своего решения.

Социальный заказ общества для лингвистов-прикладников ясен: только решение информационно-лингвистических и социально-психологических проблем банков данных позволят оптимизировать процессы проектирования, внедрения и эксплуатации БД.

ЛИТЕРАТУРА

Аграев В.А., Кобрин Р.Ю., Шульц М.М. Информационные подсистемы АСУ. - Информационные языки. - М., НС по проблеме "Кибернетика" АН СССР, 1975.

Атре Ш. Структурный подход к организации баз данных. Пер. с англ. - М., Финансы и статистика, 1983.

Бородин В.В. Процедурные грамматики. - Тезисы докладов I Всесоюзная конференция "Методы и средства обработки сложно-структурированной семантически насыщенной информации" (сентябрь, 1983). - Горький, 1983.

Будзко В.И. Предисл. к (Атре Ш., 1983).

Валькман Ю.Р. Информационная система на базе СУБД СЕДАН. - Банки данных и информационно-поисковые системы.- Киев, 1980.

Васин Ю.Г., Дмитриев С.А., Кобрин Р.Ю., Лаврова Г.К. Базовое информационно-терминологическое обеспечивание автоматизированных картографических систем. - Тезисы докладов. I Всесоюзная конференция "Методы и средства

обработки сложноструктурированной семантически насыщенной информации" (сентябрь 1983). - Горький, 1983.
Васин Ю.Г., Кустов Е.А., Миронов В.Б., Бородин В.В., Кобрин Р.Ю. Специализированный картографический автоматизированный терминал (СКАТ). - Тезисы докладов I Всесоюзная конференция "Методы и средства обработки сложноструктурированной семантически насыщенной информации (сентябрь 1983)". - Горький, 1983.
Вольфенгаген В.Э., Кузин Л.Т., Олейников В.Т. О проблеме представления информационных объектов фреймами. - Научно-техническая информация. Сер. 2, 1982, № 2.
Фольфенгаген В.Э., Кузин Л.Т., Саркисян В.Д. Реляционные методы проектирования банка данных. - Киев, Вища школа, 1979.
Глушков В., Каныгин Ю. Для всей страны. - газ. "Правда", 13 декабря 1981 г.
Дейт К. Введение в системы баз данных. - М., Наука, 1980.
Дрибас В.П. Реляционные модели баз данных. - Минск, БГУ, 1982.
РЖ "Информатика", 1981, № 1. Итоги науки и техники. Информатика - М., ВИНИТИ, 1981.
Клыков Ю.И., Горьков Л.Н. Банки данных для принятия решений. - М., Сов. радио, 1980.
Кобрин Р.Ю. Основные типы ИПС. - Вопросы построения информационно-поисковых систем. Материалы научно-технич. семинара. - Горький, Москва, ЦНИИинформ., НИИ ПМК, Горьковск. ЦНТИ, 1972.
Кобрин Р.Ю., Максимов В.Р. Место подсистемы АСНТИ в отраслевой АСУ. - Отраслевой информационный сборник ВОТ. Серия информац. - М., ЦНИИинформ., 1975 г. № 3.
Кобрин Р.Ю. О принципах терминологической работы при построении тезаурусов для ИПС. - Научно-техническая информация. Сер. 2, 1979, № 6.
Кокорева Л.В., Малашинин И.И. Проектирование банков данных. - М., Наука, 1984.
Кузин Л.Т. Банки данных для принятия решений. - В кн.: Банки информации для принятия решений. - М., МДНТП, 1976.
Мартин Дж. Организация баз данных в вычислительных системах. - М., Мир, 1980.
Овчаров Л.А., Седетков С.Н. Автоматизированные банки данных. - М., Финансы и статистика, 1982.
Олле Т.В. Предложения КОДАСИЛ по управлению базами данных. - М., Финансы и статистика, 1981.
Основные характеристики отечественных СУБД и ИПС. Под ред. Стогния А.А. - Киев, ИК АН УССР, 1980.
Пономарёва К.В., Кузьмин Л.Г., Морев В.Н. Информационное обеспечение АСУ. - М., Высшая школа, 1981.
Шенк Р. Обработка концептуальной информации. - М., Энергия, 1980.
Шуберт Л. Усиление выразительной мощности семантических сетей. - Кибернетический сборник. - М., Мир, 1979, вып.16.
Шёрс А.Л. Предисловие к: (Мартин Дж., 1980)
Ясаков Ю.В. Специализированная система хранения и обработки сложноструктурированной графической информации. - Тезисы докладов I Всесоюзная конференция "Методы и средства обработки сложноструктурированной семантически насыщенной информации (сентябрь 1983)". - Горький, 1983.
CODASYL Systems Committee. A Survey of Generalized Data Bases Management Systems. - Technical Report (May 1969).
CODASYL Systems Committee. Feature Analysis of Generalized Data Base Management Systems. - Technical Report (May 1971).

CODASYL Systems Committee. Introduction to Feature Analysis of Generalized Data Base Management Systems. - Comp. Bull, N. 4 (April 1971).
Codd E.F. Further Normalization of the Relational Model. Data Base Systems, Courant Computer Science Symposia 6, Prentice-Hall, 1972.
Codd E.F. Recent Investigation in Relational Data Base Systems. - Inf. Proc. 74, N.-H. Publ. Co., Amsterdam, 1974.
Codd E.F. Relational Completeness of Data Base Sublanguages, Data Base Systems. - Courant Computer Sci. Symposia Series, 6, Prentice-Hall, 1972.
Datenbankführer. 1. ODSN Datenbankführer. Dialogfähige Datenbanken der deutschen DJANE-Hosts DIMDI, GID und INKA Stand: März, 1983. Frankfurt/M., IDD Verl. Werner Flach, 1983.
Engles R. A Tutorial on Data Base Organizatioh. - Annual Review in Automatic Programming. Part 1.Elmsford,N.Y., Pergamon Press (July, 1972)
Heidlerová K., Chvalovský V. Ob souború dat k bankam dat. - "Mech. autom. admin.", 1983, N. 12.
Shank R. Conceptual Dependency: A Theory of Natural Language Understanding Cogn. Phye., 3, 1972.
Wells Charles H., Hopkinson Roy J. - "AlChe Symp. Ser.", 1983, No. 231.

DATA BANKS OF 80-s:
THEORY, EXPERIMENT, INCULCATION

Raphail Yu. Kobrin

S u m m a r y

The theoretical problems of the data bank projection are discussed. The data bank place among the other information systems is defined.

The author emphasizes the conceptual model for DB-projection, introduces the notion of the information and terminology support and suggests the main procedures of its projection. The technological stages of this projection are considered on the base of cartographic data.

The particular function of the computer linguistics in the data bank projection and exploitation is pointed out.

К ВОПРОСУ О ДИНАМИКЕ НАРАСТАНИЯ ОБЪЁМА СЛОВАРЯ СЛУЧАЙНОЙ ВЫБОРКИ И СВЯЗНОГО ТЕКСТА

Ю.К. Крылов

В настоящее время становится всё более очевидным, что при построении стохастических моделей, описывающих порождение текста, необходимо принимать во внимание все уровни его организации с учётом ограничений, обусловленных их взаимодействием как в системе языка в целом, так и в рамках конкретного контекста. В этом отношении не составляет исключения и проблема изучения статистических закономерностей, которым подчиняются частотные структуры лексических единиц того или иного рода.

Обычно для понимания глубинных причин ответственных за возникновение наблюдаемых ранговых распределений привлекаются различные соображения комбинаторного характера (М.В. Арапов, Ю.А. Шрейдер, 1977, 1978; Ю.К. Крылов, 1982). Вышеуказанный подход, однако, обладает тем недостатком, что в нём фактически игнорируется лингвистическая природа изучаемого объекта. Не отрицая возможностей комбинаторных методов, позволяющих интерпретировать наблюдаемые эмпирические зависимости как наиболее вероятные на множестве потенциально допустимых распределений, в предлагаемой работе делается попытка использовать для объяснения эмпирических закономерностей ряд других естественных необходимых условий.

Ещё в работах В.М. Калинина (1964, 1965) было показано, что закономерности роста объёма словаря в функции длины рассматриваемого текста однозначно определяют как лексический спектр, так и частотную структуру рангового распределения. Действительно, пусть $V_k(N)$ – объём словаря на уровне k – кратных вхождений слов в текст длиной N словоупотреблений. Другими словами, $V_k(N)$ суть число различных лексических единиц, вошедших в текст k и более раз, $V_1(N)$ – словарь в общепринятой терминологии. Очевидно, что $m_k(N)$ – число различных слов кратности k определяется разностью $m_k(N) = V_k(N) - V_{k+1}(N)$ и, таким образом, множества функциональных зависимостей $m_k(N)$ и $V_k(N)$ взаимно однозначно определяют друг друга. С другой стороны, по крайней мере для случайной выборки, справедлива система дифференциальных уравнений:

$$\frac{d\nu_i(N)}{dN} = \frac{m_1(N)}{N}$$

$$\frac{dm_k(N)}{dN} = \frac{k m_k(N) - (k+1) m_{k+1}(N)}{N}, \qquad \text{(I)}$$

которая позволяет простым дифференцированием вычислить лексический спектр, если известна функция $\nu_1(N)$. Из сказанного непосредственно вытекает, что если нам удастся понять характер и природу функциональной зависимости $\nu_1 = \nu_1(N)$, то этого понимания будет уже достаточно и для объяснения всех других количественных соотношений между частотными характеристиками структуры текста.

Необходимо однако подчеркнуть, что В.М. Калининым система уравнений (I) была получена в предположении чисто случайного характера порождения текста. При этом каждому i - му слову словаря исследуемого языка была приписана безусловная вероятность p_i его независимого появления в любой позиции текста. Естественно, что столь жёсткое допущение строго справедливо лишь при порождении "квазитекста" в виде весьма разреженной случайной выборки слов из множества различных текстов исследуемого языка. Тем не менее, как это не кажется странным на первый взгляд, модель В.М. Калинина приводит к количественным соотношениям, которые довольно хорошо выполняются и для связаных литературных текстов, о чём, в частности, свидетельствуют работы Ю.К. Орлова (1976, 1978), как известно, в качестве одного из компонентов разработанных им математических моделей использующие систему уравнений (I).

По-видимому, возможность применимости этой системы к естественным текстам связана с тем, что (I) может быть получена исходя из менее жёстких предположений о механизме порождения текста. Допустим, что элемент случайности в порождении текста обусловлен не тем, что каждое слово появляется в тексте независимо от предшествующего и будущего контекста, а лишь проявляется в том, что для любого k множество всех словоупотреблений k - кратных слов "в совокупности покрывает текст" случайным образом.

Пусть в тексте длиной $N+\Delta N$ словоупотреблений $m_k(N+\Delta N)$ слов употреблено ровно k раз. При достаточно малом ΔN естественно пренебречь возможностью "одновременного" двухкратного появления в тексте одного и того же слова, т.е. считать, что на отрезке длиной ΔN присутствуют только одноразовые слова. Тогда приращение k - разовых слов Δm_k на этом отрезке равно их числу $m_k(\Delta N)$ минус $m_{k+1}(\Delta N)$ - ко-

личество слов $(k+1)$- ой кратности, т.к. каждое из последних при присоединении к тексту отрезка ΔN увеличило кратность на единицу, т.е. перешло из класса $m_k(N)$ в класс $m_{k+1}(\Delta N)$. Суммарное число словоупотреблений слов кратности k очевидно равно $k \cdot m_k(N+\Delta N)$. Из них при случайном размещении по тексту на отрезок ΔN придётся $k \cdot m_k(N+\Delta N)\frac{\Delta N}{N+\Delta N}$ вхождений, откуда:

$$\frac{\Delta m_k}{\Delta N} = \frac{k \cdot m_k(N) - (k+1)m_{k+1}(N+\Delta N)}{N+\Delta N}. \quad (2)$$

Переходя к пределу в предположении дифференцируемости функций $m_k(N)$, сразу получим формулы В.М. Калинина

$$\frac{dm_k}{dN} = \frac{k m_k(N) - (k+1)m_{k+1}(N)}{N}. \quad (Ia)$$

Таким образом, и в случае связного текста система уравнений (I) в первом приближении может быть использована для вычисления лексического спектра, если зависимость $v_1(N)$ известна.

Переходя к обсуждению конкретного характера зависимости $v_1(N)$, прежде всего рассмотрим граничные условия при $N=1$ и $N \to \infty$. В точке $N=1$ в качестве очевидных условий выступают:

$$v_1\big|_{N=1} = m_1\big|_{N=1} = 1;\ v_k\big|_{N=1} = m_k\big|_{N=1} = 0 \text{ для } \forall\, k \geq 2. \quad (3)$$

С учётом формул (I) они приводят к

$$\frac{dv_1}{dN}\Big|_{N=1} = 1 \quad (4)$$

$$\frac{d^k v_1(N)}{dN^k} = (-1)^{k+1}\frac{k!}{N^k} m_k(N)\Big|_{N=1} = 0 \text{ для } \forall\, k \geq 2.$$

В.М. Калининым (1964) установлено, что $v_1(N)$ допускает разложение в ряд Тейлора

$$v_1(N) = v_1(N_0) + \sum_{k \geq 1} \frac{d^k v_1(N)}{dN^k} \cdot \frac{(N-N_0)^k}{k!}, \quad (5)$$

который сходится к функции $v_1(N)$ при любом N. Однако граничные условия (3) могут быть выполнены лишь для тривиальной зависимости

$$v_1(N) = N. \quad (6)$$

Полученный результат имеет большое принципиальное значение, так как показывает, что в классе аналитических функций, т.е. функций, допускающих разложение в ряд Тейлора в области естественного изменения $N \in [1, \infty[$, единственной функцией, являющейся решением задачи Коши для системы (I) при условиях (3) является линейная зависимость (6). Легко сообразить, что случаю (6) может соответствовать только словарь языка, имеющий мощность континуума, что заведомо никогда не имеет места в реальной ситуации. Отсюда следует, что никакая элементарная функция не может точно отражать зависимость $v_1(N)$ для текстов любой длины. Сказанное выше объясняет хорошо известный факт, что хотя некоторые из предложенных к настоящему времени эмпирических формул можно с большим успехом применять при интерполяции, в отношении экстраполяции эти формулы ведут себя гораздо хуже и их прогнозирующая сила оказывается гораздо слабее.

Теперь покажем, что при $N \to \infty$ функция $v_1(N)$ должна стремиться к конечному пределу. Так как словарь текста с увеличением длины последнего может только возрастать, $v_1(N)$ принадлежит классу не убывающих функций. С другой стороны, длина любой словоформы ограничена хотя бы конечным временем её произнесения. Отсюда следует, что по крайней мере в синхронном плане словарь любого языка в принципе ограничен и, таким образом, $v_1(N)$ как монотонно возрастающая ограниченная функция должна иметь конечный предел, т.е. существует

$$v_\infty = \lim_{N \to \infty} v_1(N). \qquad (7)$$

К этому же пределу будет стремиться и любая из функций $v_k(N)$, что с необходимостью для любого k приводит к условию:

$$\lim_{N \to \infty} m_k(N) = 0. \qquad (8)$$

Наблюдаемое же эмпирически для текстов любой длины увеличение числа одноразовых слов по мере возрастания N обусловлено лишь тем обстоятельством, что в выполнявшихся экспериментах длина текстов была ещё недостаточно велика для того, чтобы количество однократных слов в тексте достигло своего максимума. Однако этот эффект легко наблюдать и в текстах умеренной длины, если, например, проследить за изменением частотных характеристик словарей однобуквенных или двухбук-

венных словоформ.

По своим свойствам $\nu_1(N)$ весьма напоминает свойства интегральных функций распределения. Хорошо известно (см., например, А.Н. Колмогоров, С.В. Фомин, 1981), что последние при некоторых весьма общих условиях могут быть представлены в виде обобщённых рядов Фурье по той или иной системе ортогональных функций. Естественно допустить, что и $\nu_k(N)$ должны принадлежать классу функций, интегрируемых с квадратом на промежутке $[0, \infty[$. При этом теоретически выбор конкретной системы функций, по которому осуществляется разложение в ряд, не является существенным, так как разложение может быть осуществлено по любой системе, обладающей свойством полноты. С практической же точки зрения, однако, весьма важно, чтобы функция $\nu_1(N)$ могла быть представлена с достаточной точностью с помощью небольшого числа членов ряда. При этом, если выбор системы функций окажется удачным, можно ожидать, что коэффициенты разложения будут обладать лингвистической интерпретацией.

Чтобы остановиться на конкретной системе функций, обратимся к работе Ю.А. Тулдавы (1980). В этой работе автором исходя из соображений качественного характера была предложена эмпирическая формула

$$\nu_1(N) = N e^{-\alpha (\ln N)^{\beta}} , \qquad (9)$$

которая в весьма широком диапазоне с достаточно большой точностью описывает изменение словаря в функции длины текста как для словоформ так и для лексем. Уместно подчеркнуть, что при любых значениях параметров α и β формула (9) "автоматически" удовлетворяет важнейшим из условий (3), сформулированным выше,

$$\nu_1(N=1) = 1$$

$$\frac{d\nu_1}{dN} = e^{-\alpha (\ln N)^{\beta}} \left[1 - \alpha\beta (\ln N)^{\beta-1}\right] \Big|_{N=1} = 1, \qquad (10)$$

что, по-видимому, в значительной мере и обусловливает её положительные свойства.

Переходя к анализу (9), прежде всего покажем, что параметр β в этой формуле с необходимостью должен принимать целочисленные значения. Действительно, допустим, что (9) абсолютно точна для любого N . Пусть $N = \lambda G$ - где G неко-

торая функция, например, число графем, слогов или предложений, связанная с N с точностью до случайных флуктуаций строго линейной зависимостью. Очевидно, что при этих условиях выбор единиц измерения длины текста безразличен и мы не можем отдать предпочтение N по сравнению с G. Другими словами, формула (9) должна быть инвариантна по отношению к выбору масштаба измерения длины, т.е. должна оставаться точной для любых G и при линейной замене $N = \lambda G$. Сама же замена способна лишь привести к изменению численных значений входящих в эту формулу параметров. Покажем, что такое переопределение возможно только при целочисленных значениях β. Для того, чтобы сделать выкладки менее громоздкими, положим $\beta = 2$ (что, как показал анализ эмпирического материала, имеет место в действительности). Перепишем (9) в равносильном виде:

$$U_1 = e^{-\alpha (\ln N)^2 + \gamma \ln N + \eta}. \qquad (9а)$$

В варианте Ю.А. Тулдавы $\gamma = 1$, $\eta = 0$, так как в противном случае нарушаются условия (3). Убедимся, что замена $N = \lambda G$ снова приводит к тому же самому выражению

$$U_1 = e^{-\alpha_1 (\ln G)^2 + \gamma_1 \ln G + \eta_1}. \qquad (9б)$$

Для этого сравним степенные показатели

$$-\alpha (\ln \lambda G)^2 + \gamma (\ln \lambda G) + \eta =$$
$$= -\alpha (\ln G)^2 + (\gamma - 2\alpha \ln \lambda) \ln G + \eta + \gamma \ln \lambda - \alpha (\ln \lambda)^2,$$

равносильно (9б), если $\alpha_1 = \alpha$, $\gamma_1 = \gamma - 2\alpha \ln \lambda$,

$$\eta_1 = \eta + \gamma \ln \lambda - \alpha (\ln \lambda)^2.$$

В случае же, если β не целое, наличие множителя $\left(1 + \frac{\ln \lambda}{\ln G}\right)^\beta$ в уравнении

$$-\alpha (\ln N)^\beta + \gamma \ln N + \eta =$$
$$= -\alpha (\ln G)^\beta \left(1 + \frac{\ln \lambda}{\ln G}\right)^\beta + \gamma \ln G + \eta + \ln \lambda$$

приведёт к систематическому отклонению значений ν_1 от точной зависимости.

Итак, пусть

$$\nu_1 = e^{-\alpha (\ln N)^2 + \ln N} \qquad (9в)$$

Сразу подчеркнём, что эта функция не удовлетворяет рассмотренному выше граничному условию, т.к. при $N \to \infty$, $\nu_1 \to 0$. Однако, как отмечал сам Ю.А. Тулдава, предложенная им формула хорошо подходит и для описания динамики изменения числа однократных слов в тексте. Таким образом естественно считать, что справедливо не уравнение (9в), а выражение

$$m_1(N) = e^{-\alpha (\ln N)^2 + \ln N} . \qquad (10)$$

Для того чтобы перейти к ν_1 , используем первое уравнение системы (I)

$$\frac{d\nu_1}{dN} = \frac{1}{N} e^{-\alpha (\ln N)^2 + \ln N} , \qquad (11)$$

откуда

$$\nu_1 = \int_0^N e^{-\alpha (\ln N)^2 + \ln N} \frac{dN}{N} . \qquad (12)$$

Выполняя интегрирование, не трудно получить выражение для

$$\nu_1 = \sqrt{\frac{\pi}{\alpha}}\, e^{\frac{1}{4\alpha}} \Phi\left(\sqrt{2\alpha}\, \ln N - \frac{1}{\sqrt{2\alpha}}\right), \qquad (13)$$

где $\Phi(z)$ - известный интеграл Лапласа:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}}\, dt .$$

При $N \to \infty$

$$\lim_{N \to \infty} \nu_1 = \nu_\infty = \sqrt{\frac{\pi}{\alpha}}\, e^{\frac{1}{4\alpha}} . \qquad (14)$$

Из (I4) вытекает, что параметр α однозначно связан с потенциальным объёмом словаря изучаемого текста. Сопоставление формулы (I3) с эмпирическими данными показало, что для случайной выборки словоформ из различных текстов русского литературного языка отклонение теоретической зависимости (I3) от экспериментально наблюдавшейся траектории $\nu_1(N)$ в диапазоне из-

менения $N \sim 10^2 - 10^4$ словоупотреблений составило в среднем приблизительно 0,1 %. При этом α, вычисленное из эксперимента, привело к оценке $v_\infty \approx 5 \cdot 10^6$, что для словаря словоформ лишь в 2 – 3 раза превышает объём ССРЛЯ. Такой прогноз следует признать весьма точным, особенно если учесть, что v_∞ более чем в 1000 раз превышает объём словаря $v_1(N_{max})$ максимальной выборки, реализованной в описываемом предварительном эксперименте.

Для текстов сравнительно малой длины, когда

$$\frac{1}{\sqrt{2\alpha}} - \sqrt{2\alpha}\,\ln N \gg 1\,,$$

$v_1(N)$, даваемое формулой (13), можно разложить в ряд. Используя известное свойство интеграла Лапласа $\Phi(z) = 1 - \Phi(-z)$ и переходя к асимптотическому разложению по степеням $\frac{1}{z}$ (см., например, Б.А. Ван дер Варден, 1960, с. 23) получим:

$$\Phi(z) = 1 - \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \left(\frac{1}{z} - \frac{1}{z^3} + \frac{1\cdot 3}{z^5} - \frac{1\cdot 3\cdot 5}{z^7} + \ldots\right).$$

В этом случае

$$v_1 \simeq \frac{e^{-\alpha(\ln N)^2 + \ln N}}{1 - 2\alpha \ln N}\left[1 - \frac{2\alpha}{(1-2\alpha\ln N)^2} + \frac{1\cdot 3\cdot 4\alpha^2}{(1-2\alpha\ln N)^4} - \ldots\right] \quad (15)$$

Если же вдобавок выполняется условие $2\alpha \ln N \ll 1$, то выражение (15) переходит в (9в), что и объясняет, почему эмпирическая формула, предложенная Ю.А. Тулдавой, одинаково хорошо подходит для описания зависимостей $v_1(N)$ и $m_1(N)$.

Возвращаясь к формулам (10) и (13) отметим, что они зависят только от одного параметра α и для любых текстов с одним и тем же объёмом потенциального словаря должны приводить к идентичным зависимостям $m_1(N)$ и, как следствие, к отличающимся лишь случайными флуктуациями частотным спектрам и ранговым распределениям. Напомним, что выражения (10) и (13) не являются точными. Фактически мы отошли от намеченной ранее программы и, вместо того, чтобы строить разложение $v_1(N)$ или, что же самое, $m_1(N)$ в ряд по ортогональной системе функций, проанализировали возможности, даваемые первым членом такого разложения. Проведенные рассуждения, однако, позволили нам остановить свой выбор на конкретной системе функций. Из всего вышесказанного вытекает, что в качестве таковой следует выбрать системы полиномов по степеням $\ln N$, ортогональных с весом $Ne^{-\alpha(\ln N)^2}$ на промежутке $[0, -\infty[$.

При этом задача сводится к совместному определению как коэффициентов разложения, так и неизвестного параметра α. В силу последнего система уравнений, решение которой необходимо получить для нахождения α и неизвестных коэффициентов, становиться нелинейной, что существенно затрудняет вычисления. Кроме того, условие ортогональности оказывается мало полезным в вычислительном отношении, так как реальные наблюдения соответствуют случаю $v_1(N) \ll v_\infty$ и, в силу этого, покрывают лишь начальный участок траекторий $v_1(N)$ и $m_1(N)$. В связи с вышесказанным в данной работе снимем условие ортогональности и ограничимся рассмотрением зависимости $m_1(N)$ в виде:

$$m_1 = N e^{-\alpha (\ln N)^2} \left[1 + \sum_{j=1}^{\infty} C_j (\ln N)^j\right]. \qquad (I6)$$

Коэффициенты C_j и параметр α могут быть найдены итерационным методом. На первом шаге итераций положим $C_j = 0$ и первое приближение для α получим, например, из условия прохождения теоретической кривой через крайнюю точку наблюдавшейся траектории. Используя это значение α, приближения для коэффициентов C_j можно вычислить по методу наименьших квадратов или методу моментов. Путь дальнейших уточнений ясен. Следующее приближение для α находим, используя значения C_j, полученные на предыдущем этапе, после чего, задаваясь найденным значением α, вычисляем коэффициенты, пока не достигнем необходимой точности. После того, как значения α и C_j определены для отыскания функции $v_1(N)$, достаточно проинтегрировать уравнение (I6). Зависимости же $m_2(N)$, $m_3(N)$ и т.д. можно получить дифференцированием, используя систему уравнений (I).

В заключение отметим, что в виде

$$n_r = e^{-\alpha(\ln r)^2 - \gamma \ln r + \eta} \left[1 + \sum_{j=1}^{\infty} C_j (\ln r)^j\right] \qquad (I7)$$

найденная система функций должна служить неплохим приближением и для описания рангового распределения, т.к. при $\alpha = 0 = C_j$ формула (I7) переходит в хорошо известную форму закона Ципфа:

$$n_r = e^{-\gamma \ln r + \eta} = \frac{e^{\eta}}{e^{\gamma \ln r}} = \frac{A}{r^{\gamma}}. \qquad (I8)$$

Параметры же α и C_j должны учитывать отклонения реальных ранговых распределений от классического случая. В частности, если в (I7) пренебречь всеми слагаемыми ряда по сравнению с единицей, то получим "четвёртую модель" П.М. Алексеева (I983):

$$\ln n_r = -(\alpha \ln r + \gamma)\ln r + \eta \Rightarrow$$
$$n_r = \frac{A}{r^{\gamma + \alpha \ln r}} \qquad (I9)$$

Таким образом, оказывается, что как ранговое распределение, так и динамика роста словаря могут быть охарактеризованы одним и тем же функциональным выражением.

ЛИТЕРАТУРА

Арапов М.В., Шрейдер Ю.А. Классификации и ранговые распределения. - НТИ, Сер. 2, № II, М., I977, с. I5-2I.

Арапов М.В., Шрейдер Ю.А. Закон Ципфа и принцип диссиметрии системы. - Семиотика и информатика, № I0, М., I978, с. 74-95.

Алексеев П.М. Методика квантитативной типологии текста. Л., I983.

Ван дер Варден Б.А. Математическая статистика. М.: ИЛ, I960, с. 23.

Калинин В.М. Некоторые статистические законы математической лингвистики. - Проблемы кибернетики. Вып. II. М., I964, с. 245-255.

Калинин В.М. Функционалы, связанные с распределением Пуассона, и статистическая структура текста. - Труды Математического института им. Стеклова, том 29, М., I965, с. I82-I97.

Колмогоров А.Н., Фомин С.В. Элементы теории функций и функционального анализа. М.: Наука, I98I, с. 389-406.

Крылов Ю.К. Об одной парадигме лингвостатистических распределений. - Учёные записки ТГУ, вып. 628. Труды по лингвостатистике. Тарту, I982, с. 80-I02.

Орлов Ю.К. Обобщённый закон Ципфа-Мандельброта и частотные структуры информационных единиц различных уровней. - В кн.: Вычислительная лингвистика. М.: Наука, 1976, с. 179-202.

Орлов Ю.К. Модель частотной структуры лексики. - В кн.: Исследования в области вычислительной лингвистики и лингвостатистики. М.: Изд-во Моск. ун-та, 1978, с. 59-118.

Тулдава Ю.А. К вопросу об аналитическом выражении связи между объёмом словаря и объёмом текста. - Учёные записки ТГУ, вып. 549. Труды по лингвостатистике. Тарту, 1980, с. 113-139.

ON THE GROWTH OF VOCABULARY SIZE IN RANDOM SAMPLES AND CONNECTED TEXTS

Yuri K. Krylov

S u m m a r y

In this article the boundary conditions of a system of linear differential equations are considered which characterize the change of the lexical spectrum (word-frequency distribution) as a function of text size. It has been ascertained that one and the same analytical expression may be used for describing rank distributions as well as the change of the number of nonce words. The dependence of vocabulary size on text size may be presented by the expansion on the modified polynomials of Hermite, while in the case of a random sample the integral of the first member of the expansion gives already a sufficiently good approximation of the experimentally observed dependency.

ПРИМЕНЕНИЕ ДЗЕТА-ФУНКЦИИ РИМАНА ПРИ ПРОГНОЗИРОВАНИИ СЛОВАРНОГО СОСТАВА ПОДЪЯЗЫКА

Н.С. Манасян

Разностороннее изучение научных и технических сообщений приобрело в настоящее время большое значение. Важность этого обусловлена информационными потребностями специалистов, занятых в сфере научной и производственной деятельности. Необходимость в эффективном удовлетворении этих потребностей привела к возникновению инженерной лингвистики, которая в качестве одного из основных методов описания лингвистического объекта использует вероятностно-системный подход к изучению языка и речи, а ближайшей задачей ставит автоматизированную перера - ботку научно-технических текстов, в том числе их машинный перевод /Пиотровский/.

Применению к языку автоматических приемов обработки способствует решение ряда проблем, связанных с его формализацией. К этим проблемам относятся разного рода вероятностно-лингвис - тические, инженерно-лингвистические и информационные задачи.

В частности, теорию случайных функций и случайных процессов можно с успехом использовать при решении задач такого рода.

В работе /Манасян, 1984/ для этой цели используются частотные словари по четырем подразделениям физики: по электронике /Алексеев/, физике твердого тела /Алексеев, Каширина, Тарасова/, квантовой электронике /Манасян, 1983/ и физике элемен - тарных частиц /Алексеев, Каширина, Тарасова/. Вариационный ряд, полученный при составлении частотного словаря, рассматривается как некоторая последовательность от неслучайного аргумента $t \in \{N\}$. Каждый вариационный ряд, взятый из частотного словаря, является реализацией случайной функции $X(t)$.

Предлагаемая в работе /Манасян, 1984/ методика демонстрируется на примере однословных терминов английского подъязыка физики. Вычисляются характеристики случайной функции: математическое ожидание, среднеквадратическое отклонение и элементы

корреляционной матрицы. Каждая характеристика интерпретируется лингвистически. В частности, предполагается, что наличие функциональной зависимости между сечениями случайной последовательности предполагает существование семантических связей между сечениями случайной последовательности.

Было установлено, что математическое ожидание, так же, как реализация случайной функции, подчиняются гиперболическому закону. Этому же закону подчиняется среднеквадратическое отклонение случайной последовательности. Этот факт, а также определение верхних и нижних границ доверительного интервала математического ожидания дают возможность вычислить теоретическое среднеожидаемое количество однословных терминов в английском подъязыке физики с помощью дзета-функции Римана.

Кроме того, с помощью этой же функции на основании формул приближающих функций вычисляется среднеожидаемое количество однословных терминов в электронике, физике твердого тела, квантовой электронике и физике элементарных частиц.

Рассмотрим это подробнее.

Каждый частотный словарь рассматривается как реализация некоторой случайной последовательности /или некоторого массового процесса/, как некоторая пробная группа /выборка/, представляющая только один из возможных результатов, которые могли бы получиться при продолжении наблюдений над генеральной совокупностью /данного подъязыка/. При этом приходится помнить, что выводы и оценки, основанные на нашем ограниченном лингвистическом материале наблюдений, отражают случайный состав нашей пробной группы и потому должны считаться приближенными оценками вероятностного характера. Тем не менее, теория вероятностей во многих случаях указывает, как наилучшим способом использовать имеющуюся информацию в частотном словаре для получения по возможности более точных и надежных характеристик с указанием степени надежности наших выводов, которые возникают вследствие ограниченности наших наблюдений. Без указания степени надежности выводов заключения не будут иметь достаточной научной ценности. Один из первостепенных вопросов, которые возникают при этом,

вопрос о представительности /репрезентативности/, подробно рассмотрен в /Пиотровский, Бектаев, Пиотровская/. Второй основной вопрос - это выявление оценок параметров, т.е. указание границ изменения этих параметров.

Любая оценка $\tilde{a}$ параметра a, вычисленная на основе нашего лингвистического материала, является функцией значений случайной величины

$$\tilde{a} \approx \tilde{a}(x_1(t), x_2(t), \ldots, x_n(t)). \qquad /1/$$

Как известно из теории вероятностей, к оценке /I/ предъявляется ряд требований /см., например, Смирнов, Дунин-Барковский, гл.УI §I/.

Отметим, что при лингвостатистических исследованиях не всегда возможно, чтобы оценки удовлетворяли всем этим требованиям. Иногда для простоты расчетов применяются незначительно смещенные оценки. Но при этом, с самого начала к этим оценкам нужно отнестись критически, рассматривая их со всех со всех точек зрения.

В большинстве случаев для оценки математического ожидания пользуются формулой

$$\tilde{m}_x(t_2) = \frac{\sum_{i=1}^{n} x_i(t_2)}{n}. \qquad /2/$$

Эта оценка является состоятельной, что непосредственно вытекает из закона больших чисел, и несмещенной. Что касается дисперсии этой оценки

/3/

то ее эффективность или неэффективность зависит от закона распределения сечения случайной функции. Доказывается, что если этот закон нормальный, то эта дисперсия является минимально возможной, т.е. оценка $\tilde{m}$ является минимально эффективной. Доказывается также /Вентцель, с.3I5/, что оценку дисперсии можно

вычислить по формуле

$$\widetilde{\mathcal{D}}_x(t_2)=\frac{n}{n-1}\left\{\frac{\sum_{i=1}^{n}[x_i(t_2)}{n}-[m_x(t_2)]^2\right\} \qquad /4/$$

Оценки /2/ и /4/ являются точечными, так как выражаются одним числом. При статистической обработке ограниченного числа наблюдений важно знать еще точность этой оценки и надежность. Для этого вводятся понятия доверительного интервала и доверительной вероятности. Объясним эти понятия на примере математического ожидания случайной последовательности однословных терминов.

Пусть для математического ожидания $m_x(t)$ получена несмещенная оценка $\widetilde{m}_x(t)$. Для того, чтобы получить представление о точности и надежности этой оценки, можно для каждого β, близкого к единице, указать такое положительное число ε_β, что вероятность

$$p(|\widetilde{m}-m|<\varepsilon_\beta)=\beta. \qquad /5/$$

Ясно, что для данного β чем меньше ε_β, тем точнее оценка $\widetilde{m}$. Из /5/ следует, что

$$p(\widetilde{m}-\varepsilon_\beta<m<\widetilde{m}+\varepsilon_\beta)=\beta. \qquad /6/$$

Из /6/ следует, что с вероятностью β неизвестное значение математического ожидания попадает в интервал

$$I_\beta(\widetilde{m}-\varepsilon_\beta,\ \widetilde{m}+\varepsilon_\beta). \qquad /7/$$

Такой интервал называется доверительным интервалом, а вероятность β - доверительной вероятностью. Числа $m_1=\widetilde{m}-\varepsilon_\beta$ и $m_2=\widetilde{m}+\varepsilon_\beta$ называются доверительными границами. Значения параметра m, которые выходят из доверительного интервала нужно считать противоречащими нашим опытным данным, а те, которые внутри I_β - совместимыми с этими данными.

При этом считается, что событие с вероятностью β является практически почти достоверным, а $\alpha=1-\beta$ практически

почти невозможным событием /при достаточно большой вероятности β /. Покажем, каким образом можно построить границы доверительного интервала для математического ожидания случайной последовательности однословных терминов. Если бы был известен закон распределения величины $m_x(t)$, то мы нашли бы такое значение ε_β, чтобы выполнялось соотношение /5/. Но этот закон неизвестен, он зависит от закона распределения $x(t)$, то есть от неизвестных параметров.

Установление закона распределения сечения случайной последоватедьности затруднено тем, что в наличии мы имеем довольно малое число реализаций, т.е. число частотных словарей. Получение достаточного числа реализаций связано с большими расходами и огромным трудом целого коллектива. Поэтому в настоящем исследовании пришлось искать способ, при помощи которого можно сделать более или менее надежные статистические выводы, довольствуясь только реализациями, которые имелись в налиции.

С этой целью были взяты 48 серий по I тыс. словоупотреблений каждая, в которые вошли тексты по физике, включая и те подъязыки, по которым были составлены частотные словари, рассматриваемые в данном исследовании. Объединение двух таких серий дало 24 серии по 2 тыс. словоупотреблений, которые и послужили основой для анализа вариационных рядов при проверке гипотезы о нормальном распределении. Гипотеза о нормальности распределений сечений случайной последовательности подтвердилась. Тогда согласно центральной предельной теореме оценка математического ожидания $\tilde{m}_x(t)$ как сумма n независимых одинаково распределенных величин при достаточно большом n тоже распределена нормально. Характеристиками этого закона можно принять $m_x(t)$ и $\sigma_x(t)/n$; см. /Вентцель, гл.I3 §3/, где доказывается, что доверительный интервал для математического ожидания равен

$$I_\beta =]\tilde{m}_x(t) - \lambda_\beta \sigma_{\tilde{m}}(t), \tilde{m}_x(t) + \lambda_\beta \sigma_{\tilde{m}}(t)[, \qquad /8/$$

где $\tilde{m}_x(t)$ - оценка математического ожидания $x(t)$, а $\sigma_{\tilde{m}}(t)$ - среднеквадратическое отклонение оценки математического ожидания, которое на основании /2/ определяется формулой

$$\sigma_{\tilde{m}}(t) = \sqrt{\frac{\mathcal{D}_x(t)}{n}} \,. \qquad /9/$$

Что касается λ_β, то оно является некоторой функцией от доверительной вероятности β, для которой составлена специальная таблица /см. Вентцель, с.301/.

Лингвостатистический смысл λ_β для случайной последовательности однословных терминов в случае нормального распределения по ее ординатам можно видеть в следующем: это число средних квадратических отклонений, которое нужно отложить от центра рассеивания для того, чтобы вероятность попадания в полученный интервал /8/ была равна β. Из формулы /8/ найдем нижнюю и верхнюю границы доверительного интервала

$$\begin{cases} m_1(t) = \tilde{m}_x(t) - \lambda_\beta \sigma_{\tilde{m}}(t) \\ m_2(t) = \tilde{m}_x(t) + \lambda_\beta \sigma_{\tilde{m}}(t) \,. \end{cases} \qquad /10/$$

Ясно, что при данной доверительной вероятности β границы доверительной вероятности являются функциями от частоты t.

Теперь для каждой частоты найдем доверительный интервал, и по этим доверительным интервалам построим доверительную область, которая покажет, что с вероятностью β все реализации случайной последовательности однословных терминов не выходят за пределы этой области. Результаты вычислений занесены в таблицу, которая здесь из-за дефицита места не приводится./Это относится и к другим вычислениям, имеющим табличную форму/.

Для того, чтобы воспользоваться формулами /10/ в целях построения теоретических доверительных интервалов, функцию нужно аппроксимировать так, чтобы она наилучшим образом описывала $\sigma_{\tilde{m}}(t)$. Предполагая, что распределение этой функции относится к гиперболическому типу[1], графически[2], т.е. способом "натянутой нити", находим, что

---

[1] В пользу такого распределения говорит вид кривой $\sigma_{\tilde{m}}$ /см. Рис.1/.

[2] Так как $\sigma_{\tilde{m}}$ вычисляется по грубой формуле /9/, то точные вычисления параметров функции $\sigma_{\tilde{m}}(t)$ были бы некорректными.

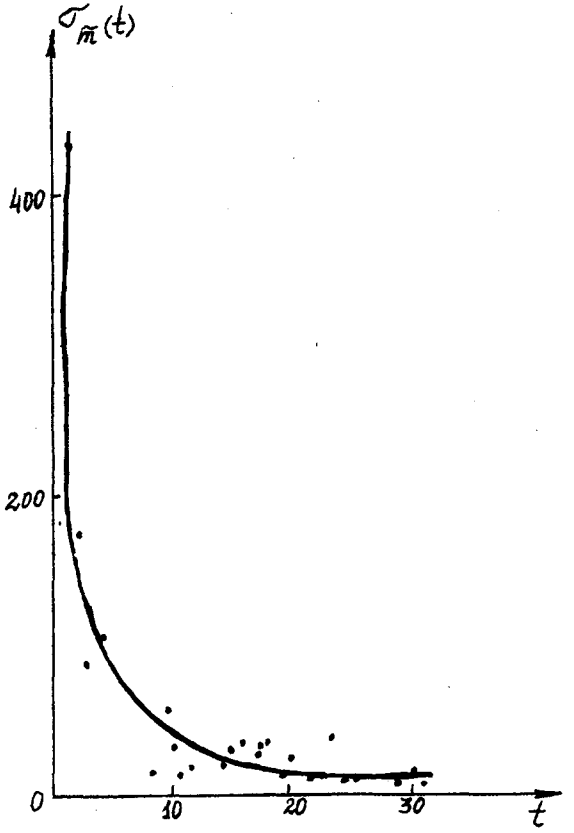

Рис.I. Распределение среднеквадратичного отклонения математического ожидания случайной последовательности одноименных терминов.

$$\sigma_{\tilde{m}}(t) = \frac{110}{t^{1,20}} . \qquad /11/$$

Эта формула довольно хорошо аппроксимирует ряд для $\sigma_{\tilde{m}}$.

Если учесть, что

$$\tilde{m}_x(t) = 1516 \cdot t^{-1,37} . \qquad /12/$$

/см. Манасян, 1984/, и подставляя /11/ и /12/ в /10/, получим следующие формулы для доверительных границ

$$\begin{cases} m_1(t) = 1516 \cdot t^{-1,37} - 110\lambda_\beta \cdot t^{-1,20} \\ m_2(t) = 1516 \cdot t^{-1,37} + 110\lambda_\beta \cdot t^{-1,20} \end{cases} \qquad /13/$$

На их основании можно судить о генеральной совокупности, если предположить, что она практически бесконечна. Как было указано выше, за генеральную совокупность принимется общее количество однословных терминов в английском подъязыке физики. Формулы/13/ дают возможность вычислить теоретическое значение числа однословных терминов в рассматриваемой генеральной совокупности. Если это число обозначить через Q , то будем иметь

$$Q = \sum_{t=1}^{\infty} (1516 \cdot t^{-1,37} \pm 110\lambda_\beta \cdot t^{-1,20}) = .$$

$$= 1516 \sum_{t=1}^{\infty} \frac{1}{t^{1,37}} \pm 110\lambda_\beta \sum_{t=1}^{\infty} \frac{1}{t^{1,20}} . \qquad /14/$$

Как видно из /14/, вычисление числа Q приводится к суммированию рядов вида

$$a \sum_{t=1}^{\infty} \frac{1}{t^S} ,$$

где a и S - известные параметры. Суммирование рядов такого типа непосредственно связано с дзета-функцией Римана.

Итак, для того, чтобы получить количество всех однословных терминов, нужно вычислить сумму ряда

$$\sum_{t=1}^{\infty} \frac{a}{t^{1+\delta}} = a \sum_{t=1}^{\infty} \frac{1}{t^{s}}. \qquad /I5/$$

Для вычисления суммы ряда /I5/ рассмотрим ряд

$$\zeta_R(s) = \sum_{t=1}^{\infty} \frac{1}{t^{s}}, \qquad /I6/$$

где $s = I + \delta$. Здесь $t \in N$, $\delta > 0$ /в нашем лингвистическом приложении δ и t вещественны/. Ряд /I5/ называется дзета-функцией Римана. Эта функция была известна еще Эйлеру, но наиболее значительные ее свойства были открыты Риманом в 70-80 гг. прошлого века. Полная теория этой функции изложена в фундаментальном труде Е.К.Тичмарша /Тичмарш/ и в /Уитекнер, Ватсон,т.2, с.58-80/. Вкратце свойства дзета-функции Римана изложены в /Янке, Эмде, с.I68-I79/.

Поскольку в нашем приложении $s = I + \delta$ $/\delta > 0/$ принимает только вещественные значения, ниже будут даны некоторые свойства этой функции именно для этого случая, причем будут приведены те свойства, которые найдут приложения в данной работе.

I. Дзета-функция Римана является убывающей, т.е. если $s_1 > s_2$, то

$$\zeta_R(s_1) < \zeta_R(s_2). \qquad /I7/$$

Это свойство надо учитывать при нахождении значений дзета-функции по ее таблицам при помощи интерполяции.

2. Значения дзета-функции Римана выражаются через суммы сходящихся рядов с положительными членами, то есть ряд /I5/сходится при условии, что $s > I$. Это следует из интегрального признака Коши /см. Фихтенгольц, с.I6-I7/ и из равенства /I5/ и условия $\delta > 0$. Ниже в таблице I приводится та часть таблицы дзета-функции, которая пригодится в нашем приложении и заимств

вована из /Янке, Эмде, с.372 и сл./, где также приведены правила пользования этой таблицей.

Таблица I

S		0	I	2	3	4	5	6
I	+0.00Ix	∞	I0584	5592	3932	3I06	26I2	2286

3. При $S = I$ ряд /I6/ расходится, так как он превращается в известный гармонический ряд /см.Фихтенгольц, с.I6/. В лингвистическом отношении представляет определенный интерес поведение дзета-функции при $S \to I$. Доказывается /см. Уитеккер, Ватсон, т.2, с.66/, что

$$\lim_{S \to 1} \left\{ \zeta(S) - \frac{1}{S-1} \right\} = \gamma , \qquad /I8/$$

где $\gamma = 0.5772I57...$ - постоянная Эйлера, равная

$$\gamma = \lim_{n \to \infty} \left(1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{m} - \ln m\right). \qquad /I9/$$

Имея в виду /I5/, /I8/ переписываем как

$$\lim_{\delta \to 0} \left\{ \zeta(1+\delta) - \frac{1}{\delta} \right\} = \gamma . \qquad /20/$$

По определению предела из /20/ найдем

$$\zeta(1+\delta) = \gamma + \frac{1}{\delta} + \alpha(\delta), \qquad /2I/$$

где $\alpha(\delta) \to 0$ при $\delta \to 0$. Тогда из /2I/ можно написать следующее приближенное равенство

$$\zeta(1+\delta) \approx \gamma + \frac{1}{\delta} . \qquad /22/$$

4. Имеет место следующая оценка дзета-функции Римана

$$\frac{1}{\delta} < \zeta(1+\delta) < 1 + \frac{1}{\delta} . \qquad /22a/$$

Итак, используя дзета-функцию Римана можно вычислить теоретически ожидаемое число терминов в среднем в английском подъязыке физики.

Формулу /I3/ в обозначениях дзета-функции можно переписать как

$$Q = 1516 \cdot \zeta_{\alpha}(1.37) \pm 110 \lambda_{\beta} \zeta_{\alpha}(1.20) \qquad /23/$$

Значения дзета-функции, входящие в правую часть /23/, найдем из таблицы I при помощи интерполирования, имея в виду свойство I рассматриваемой функции

$$\zeta_{\alpha} \ /1.37/ = 3{,}354$$
$$\zeta_{\alpha} \ /1.20/ = 5{,}592.$$

Тогда, подставляя эти значения в /23/, получим, что

$$Q = 5085 \pm 615 \cdot \lambda_{\beta}. \qquad /24/$$

Чтобы вычислить Q, остается выбрать только ту или иную доверительную вероятность. При 80%-ой доверительной вероятности / $\beta = 0.8$/ по таблице из /Вентцель, с.301/ находим $\lambda_{\beta} = 1.282$. Тогда

$$Q = 5085 \pm 788$$
$$4297 < Q < 6095.$$

Совершенно аналогично вычисляется ожидаемое количество однословных терминов при доверительной вероятности $\beta = 0.9$

$$4075 < Q < 6095.$$

Используя вышеизложенную методику, вычислим теоретически ожидаемой количество однословных терминов, например, для каждого из физических подъязыков, приведенных в /Манасян, 1984/

$$Q_{э} = \sum_{t=1}^{\infty} x_1(t) = 1346 \cdot \zeta_{\alpha}(1.31) = 5182$$

$$Q_{фтт} = 1510 \cdot \zeta_{\alpha}(1.34) = 5439.$$

$$Q_{кэ} = 2235 \cdot \zeta(1.54) = 6128$$

$$Q_{физ} = 1741 \cdot \zeta(1.45) = 4978$$

Дзета-функция может иметь также несколько другое применение. При построении аппроксимирующих функций для реализаций случайной последовательности мы видим, что параметр $a_i \to S$ меняется от I.2I до I.54. Не исключена возможность того, что могут быть такие реализации, при которых $S \to I$, оставаясь больше единицы. Вариационный ряд с такими характеристиками может быть получен только по частотным словарям, составленным по текстам с наибольшим возможным количеством терминов /назовем их "перенасыщенными" терминами текстами/[1]. Для подсчета количества терминов в таких текстах можно использовать асимптотическое равенство /22/. Для этого обе части этого равенства умножим на a, где a - число терминов с частотой, равной единице, и получим

$$a\zeta(1+\delta) \approx a\left(\gamma + \frac{1}{\delta}\right)$$

или

$$Q_{max} \approx a\left(\gamma + \frac{1}{\delta}\right).$$

Например, если a = I500, δ = 0.I, то Q_{max} = I5866 /терминов/.

Еще одно применение дзета-функции состоит в следующем. Для данного δ неравенство /22а/ дает возможность оценить число однословных терминов для данного частотного словаря. На самом деле, умножим обе части неравенства /22а/ на a, где a - число однословных терминов с частотой I

$$\frac{a}{\delta} < a\zeta(1+\delta) < \left(1 + \frac{1}{\delta}\right)a,$$

---

[1] К подобным текстам могут относиться, например, некоторые разделы патентов.

но $aG_s/1+\delta/ = Q$, где Q - число всех однословных терминов в генеральной совокупности. Тогда

$$\frac{a}{\delta} < Q < (1+\frac{1}{\delta})a.$$

Такова невероятностная оценка числа всех однословных терминов. Если, например, a = 1600, а δ = 0.4, то получим

$$4000 < Q < 5600$$

Таким образом, применение дзета-функции позволяет определить теоретически ожидаемое число терминов в специальных подъязыках, что является важным для решения ряда прикладных задач лингвистического плана. К таким задачам относится, в частности, прогнозирование достаточных объемов словаря в автоматизированных информационных системах.

Следует отметить, что оценки генеральной совокупности при помощи дзета-функции Римана могут иметь как вероятностный, так и невероятностный характер, в зависимости от применяемых формул.

Значение величины δ может служить диагносцирующим инструментом, информирующем о насыщенности текста терминами.

ЛИТЕРАТУРА

Алексеев П.М. Частотный словарь английского подъязыка электроники. - М.: Воениздат, 1971. - 302 с.

Алексеев П.М., Каширина М.Е., Тарасова Е.М. Частотный англо-русский физический словарь-минимум. - М.: Воениздат, 1980. - 288 с.

Вентцель Е.С. Теория вероятностей. М.: Наука, 1969. - 366с.

Манасян Н.С. К вопросу о применении теории случайных функций при изучении квантитативных особенностей лингвистических систем. - Учен. зап. Тартуского ун-та, вып. 689. Труды по лингвостатистике, Тарту, 1984, с. 78-92.

Манасян Н.С. Частотный англо-русский словарь-минимум по квантовым генераторам. - М.: Воениздат, 1988. - 272 с.

Пиотровский Р.Г. Инженерная лингвистика и теория языка.-Л.: Наука, 1979. - 112с.

Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А. Математическая лингвистика.-Л.: Высш. школа, 1977. - 383 с.

Смирнов М.В., Дунин-Барковский И.В. Курс теории вероятностей и математической статистики для технических приложений. М.: Наука, 1965. - 511с.

Тичмарш Е.К. Дзета-функция Римана. М., 1947. - 156 с.

Уиттекер Э.Т., Ватсон Дж.И. Курс современного анализа.-М.: Физматгиз, 1963. - 515 с.

Фихтенгольц Г.М. Основы математического анализа. Том I.-М.: Наука, 1964. - 440 с.

Янке Е., Эмде Ф. Таблицы функций с формулами и кривыми.-М.: Физматгиз, 1959. - 420 с.

THE APPLICATION OF RIEMANN'S DZETA-FUNCTION IN VOCABULARY STRUCTURE PREDICTION

Narineh Manasyan

S u m m a r y

The description of the methodics and results of the application of the above mentioned function is presented in the article. For this purpose the data of four frequency dictionaries are used.

СТАТИСТИЧЕСКИЕ ПАРАМЕТРЫ СЛОВООБРАЗОВАНИЯ В БОЛГАРСКОМ ЯЗЫКЕ XVIII В.

М.П. Рускова

Настоящая статья представляет собой опыт характеристики различных статистических параметров словообразования имен существительных в болгарском языке XVIII в. В связи с отсутствием работ по статистическому исследованию словообразования имен существительных ниже были использованы некоторые параметры и типы распределений, применяемые при статистическом анализе лексики и морфологии (Гиро П., 1954; Тенителова М., 1966; Бектаев К., Лукьяненков К., 1971, Орлов Ю., 1971; Ненитой В., 1973; Каширина М., 1974; Тулдава Ю., 1976; Надарейшвили И., Орлов Ю., 1978; Тенителова М., 1980; Тулдава Ю., 1980; Якубайтис Т., 1981; Петрушевич М., 1983).

Статистический анализ языковых фактов памятников письменности имеет свою специфику. Прежде всего нельзя не отметить, что многие памятники письменности часто не очень большого объема. Даже, если привлечь материал из нескольких памятников, которые дают возможность охарактеризовать достаточно полно какой-либо жанр или отдельные языковые явления, связанные с тем или иным центром письменности или эпохой, то объем всей выборочной совокупности будет относительно невелик. Другая особенность – это большое число низкочастотных фактов, которые должны быть объектом особого внимания. Ниже значительную роль при статистической характеристике фактов играет морфемный и словообразовательный анализ рассматриваемых существительных. Этот анализ проводится на базе созданных ранее частотных словарей (Рускова 1978, 1983). Для каждого из пяти обследованных памятников (см. список источников) был создан частотный словарь знаменательных слов, при объеме выборки в 24000 словоупотреблений, что дало в итоге общую выборочную совокупность с абсолютной накопленной частотой (F^*) в 120000 словоупотреблений. Существительное считалось членимым и соответственно выделялась та или иная морфема, только если в обследованном массиве текстов была зафиксирована соотносимая по значению и по форме производящая основа.

Проведенный анализ дал возможность выявить как число всех разных знаменательных слов (L = 5250) во всех памятниках, так и число разных имен существительных во всем массиве: N' = 1719. Доля последних составляет 32,74 % по отно-

шению ко всем разным словам. При этом для производных существительных L = 934, а для непроизводных - L = 784. Таким образом, доля производных среди всех различных существительных составляет 54,33 %, что на 10,47 % превышает долю непроизводных существительных. По нашим данным доля непроизводных существительных составляет 43,86 %.

Производные существительные образованы семью способами: суффиксальным, префиксально-суффиксальным, префиксальным, префиксально-флективным, сложением основ, сложением с суффиксацией, а также способом нулевой суффиксации. Распределение числа различных существительных и их абсолютная накопленная частота (F*) по отдельным способам показаны в таблице № I.

Таблица № I.

число разных слов и их абс. накоп. частота / способы словообразования	L	F*
суффиксальный	739	5209
префиксально-суффиксальный	I5	57
префиксальный	22	I84
префиксально-флективный	2I	97
нулевая суффиксация	53	I84
сложение основ	9	43
сложение с суффиксацией	75	I64

Факты, приведенные в таблице I показывают, что как по числу различных слов, так и по их абсолютной накопленной частоте первое место принадлежит производным существительным, образованным суффиксальным способом. Суффиксальные существительные занимают более 2/3 из всех производных существительных и более I/3 из всех различных существительных. Доля суффиксальных существительных составляет 79,I2 % среди производных существительных и 42,99 % среди всех существительных. Суффиксальному способу противопоставлены все остальные способы. В порядке убывания числа разных слов, образованных несуффиксальным путем остальные способы располагаются следующим образом: сложение с суффиксацией, нулевая суффиксация, префиксальный, префиксально-флективный, префиксально-суффиксальный, сложение основ. По убыванию абсо-

лютной накопленной частоты существительных способы распределяются так: нулевая суффиксация и префиксальный, сложение с суффиксацией, префиксально-суффиксальный, сложение основ.

Таким образом данные таблицы № I говорят о том, что по своей продуктивности такие способы образования существительных, как сложение с суффиксацией и нулевая суффиксация занимают второе и третье место после суффиксального. Распределение существительных, образованных несуффиксальным способом в порядке убывания их доли среди всех производных существительных следующее:

I) доля существительных, образованных путем сложения основ с суффиксацией 8,02 %;
2) доля существительных, образованных путем нулевой суффиксации 5,67 %;
3) доля префиксальных существительных 2,35 %;
4) доля префиксально-флективных существительных 2,24 %;
5) доля префиксально-суффиксальных существительных I,60 %;
6) доля существительных, образованных путем сложения основ 0,97 %.

Приведенные факты говорят о том, что доля суффиксальных существительных среди всех производных существительных (79,I2%) резко противопоставлена доля существительных, образованных другими способами (20,88 %). Из 20,88 % наибольшую часть занимает доля существительных, образованных способом сложения с суффиксацией (8,02 %), за которым следует доля существительных, образованных способом нулевой суффиксации (5,67 %). Приблизительно одинакова доля существительных, образованных префиксальным и префиксально-флективным способами (2,35 % и 2,24 %). Только доля существительных, образованных сложением основ меньше единицы (0,97 %), в чем она противопоставлена долевому распределению всех остальных способов.

Различия между отдельными способами словообразования хорошо видны при сравнении их параметрических индексов. Рассмотрим эти индексы.

I Индексы способа суффиксации показывают отношение числа всех разных существительных, образованных суффиксальным способом (S1) к: I) числу всех различных знаменательных слов (L); 2) числу всех различных существительных (N'); 3) числу всех производных существительных (N'_1). В наших материалах индексы суффиксации соответственно имеют следующие значения: I) $\frac{SL}{L} = 0,I4$; 2) $\frac{SL}{N'} = 0,42$; 3) $\frac{S1}{N'_1} = 0,79$.

II Индексы сложения с суффиксацией показывают отношение числа всех разных существительных, образованных способом сложения с суффиксацией (RRS) к: I) числу всех различных знаменательных слов (L); 2) числу всех различных существительных (N'); 3) числу всех производных существительных (N_1'). Рассматриваемые индексы соответственно имеют значения: I) $\frac{RRS}{L}$ = 0,0I; 2) $\frac{RRS}{N'}$ = 0,04; 3) $\frac{RRS}{N_1'}$ = 0,08.

III Индексы нулевой суффиксации показывают отношение числа всех разных существительных, образованных способом нулевой суффиксации (ØS) к: I) числу всех различных знаменательных слов (L); 2) числу всех различных существительных (N'); 3) числу всех производных существительных (N_1'). Индексы нулевой суффиксации соответственно имеют значения: I) $\frac{ØS}{L}$ = 0,0I; 2) $\frac{ØS}{N'}$ = 0,03; 3) $\frac{ØS}{N_1'}$ = 0,05.

IУ Индексы префиксации показывают отношение числа всех разных существительных, образованных префиксальным способом (P) к: I) числу всех различных знаменательных слов (L); 2) числу всех различных существительных (N'); 3) числу всех производных существительных (N_1'). Индексы префиксации отмечены со значениями: I) $\frac{P}{L}$ = 0,004; 2) $\frac{P}{N'}$ = 0,0I; 3) $\frac{P}{N_1'}$ = 0,02.

У Индексы префиксальной флективности показывают отношение числа разных существительных, образованных префиксально-флективным способом (PF) к: I) числу всех различных знаменательных слов (L); 2) числу всех различных существительных (N'); 3) числу всех производных существительных (N_1'). Приведенные индексы имеют соответственно значения: I) $\frac{PF}{L}$ = 0,003; 2) $\frac{PF}{N'}$ = 0,0I; 3) $\frac{PF}{N_1'}$ = 0,02.

УI Индексы префиксальной суффиксации показывают отношение числа разных существительных, образованных префиксально-суффиксальным способом (PS) к: I) числу всех различных знаменательных слов (L); 2) числу всех различных существительных (N'); 3) числу всех производных существительных (N_1'). Индексы префиксальной суффиксации имеют соответственно значения: I) $\frac{PS}{L}$ = 0,002; 2) $\frac{PS}{N'}$ = 0,008; 3) $\frac{PS}{N_1'}$ = 0,0I.

УII Индексы основосложения показывают отношение числа разных существительных, образованных способом сложения основ (RR) к: I) числу всех различных знаменательных слов (L); 2) числу всех различных существительных (N'); 3) числу всех производных существительных (N_1'). По имеющимся дачным индексы основосложения соответственно имеют значе-

ния: I) $\frac{RR}{L}$ = 0,001; 2) $\frac{RR}{N}$ = 0,005; 3) $\frac{RR}{N}$ = 0,009.

Сопоставление этих индексов показывает, что во всех случаях соотношение между индексами остается одинаковым – наибольшее значение имеют индексы суффиксации и наименьшее – индексы сложения основ.

Поскольку больше всего слов образовано суффиксальным способом рассмотрим также и некоторые другие параметры, связанные с ним. Общее число различных суффиксов, при помощи которых образуются отдельные суффиксальные существительные во всех обследованных памятниках – 88. Из них 52 суффикса присоединяются к основе глагола, 48 – к основе существительного, 2I – к основе прилагательного и только два – к основе числительного. К основам всех четырех различных частей речи присоединяется только один суффикс ица[I]; к трем основам – 9 суффиксов, из которых один – ар – к основам глагола, существительного, числительного и 8 – ак, ец[I], ме[I], ина[I], ник, ня, ствие, ство – к основам глагола, существительного и прилагательного; к двум основам присоединяются I4 суффиксов, из которых 8 – не[2], ика, ин, ище, ия[I], на[I], ница, ца – к основам глагола и существительного, 3 – да, ество, ост – к основам глагола и прилагательного, 3 – ик, иня, ота – к основам существительного и прилагательного; 64 суффикса присоединяются только к одной основе: 3I – только к основе глагола, 27 – только к основе существительного и 6 – только к основе прилагательного.

Ниже приводится частотный список всех суффиксов суффиксальных существительных, расположенных в порядке убывания их абсолютных накопленных частот по отношению ко всей выборочной совокупности I20000 словоупотреблений.

i	суффикс	F*	i	суффикс	F*	i	суффикс	F*
I.	ение	499	30.	че	37	59.	лище	6
2.	ица^{I}	36I	3I.	анин	35	60.	лник	6
3.	ина^{I}	345	32.	ест	33	6I.	овица	5
4.	ин	309	33.	ие^{2}	32	62.	ец^{2}	4
5.	ство	287	34.	ло	32	63.	ица^{2}	4
6.	ота	272	35.	ен	3I	64.	овище	4
7.	янин	268	36.	ен	3I	65.	ека	3
8.	ние	26I	37.	ствие	29	66.	ия^{2}	3
9.	ост	I89	38.	да	28	67.	лка	3
I0.	ец^{I}	I85	39.	ва	25	68.	лница	3
II.	ба	I6I	40.	ист	2I	69.	ца	3
I2.	ик	I56	4I.	тир	I9	70.	ек	2
I3.	ник	I43	42.	ия^{I}	I8	7I.	ел	2
I4.	ика	I42	43.	ница	I7	72.	ечка	2
I5.	ар	III	44.	ие^{I}	I5	73.	ице	2
I6.	иня	I08	45.	ина^{2}	I4	74.	иче	2
I7.	ах	89	46.	енец	I4	75.	ка^{2}	2
I8.	ък	86	47.	янство	I3	76.	лец	2
I9.	не	76	48.	мо	I3	77.	ух	2
20.	ов	72	49.	ня	I2	78.	ава	I
2I.	тел	69	50.	овство	I2	79.	еница	I
22.	ество	62	5I.	ач	9	80.	енка	I
23.	тие	59	52.	ене	9	8I.	ествие	I
24.	тва	56	53.	аче	8	82.	еч	I
25.	ка^{I}	53	54.	енце	8	83.	еш	I
26.	еник	47	55.	янка	8	84.	ичка	I
27.	джия	45	56.	тък	8	85.	унка	I
28.	ище	44	57.	нина	7	86.	уч	I
29.	ак	4I	58.	иво	6	87.	це	I
						88.	щина	I

В приведенном частотном списке выделяются три зоны высокочастотная (i = I–I6), где частота суффиксов выше I00 среднечастотная (i = I7–40) с частотой ниже I00, но не менее 20 и низкочастотная (i = 4I–88) с частотой суффиксов ниже 20. Итак высокочастотная зона представлена I6 суффиксами, среднечастотная – 24 и низкочастотная – 48 суффиксами. При распределении суффиксов, встречающихся только с одним словом низкочастотная зона резко противопоставлена остальным двум. Среди суффиксов высокочастотной зоны только один суффикс – янин – отмечен в одном слове, а в среднечастотной зоне –

только 4 суффикса – ах, ов, ест, ист. В то же время в низкочастотной зоне в одном слове встретилось 30 суффиксов, что составляет 64 % из всех различных низкочастотных суффиксов. Среди суффиксов низкочастотной зоны 38 (т.е. 78 %) имеют частоту ниже I0. При этом I9 суффиксов отмечены с частотой 2 и I.

Ниже рассмотрим ряд индексов, в которых отражено различие между существительными, образованными суффиксальным способом в зависимости от тех основ, с которыми они связаны. Сюда относятся следующие индексы.

I Индексы, выражающие отношение числа всех разных существительных, образованных суффиксальным способом от производящей глагольной основы (Sl_V) к:

I) числу всех различных существительных (N'); 2) числу всех разных производных существительных (N'_1); 3) числу всех разных суффиксальных существительных (Sl). Рассматриваемые индексы имеют значения: I) $\frac{Sl_V}{N'} = 0,23$; 2) $\frac{Sl_V}{N'_1} = 0,44$; 3) $\frac{Sl_V}{Sl} = 0,55$.

II Индексы, выражающие отношение числа всех разных существительных, образованных суффиксальным способом от основы имени существительного (Sl_N) к: I) числу всех различных существительных (N'); 2) числу всех разных производных существительных (N'_1); 3) числу всех разных суффиксальных существительных (Sl). Эти индексы имеют соответственно значения: I) $\frac{Sl_N}{N'} = 0,I2$; 2) $\frac{Sl_N}{N'_1} = 0,23$; 3) $\frac{Sl_N}{Sl} = 0.29$.

III Индексы, выражающие отношение числа всех разных существительных, образованных суффиксальным способом от основы имени прилагательного (Sl_A) к: I) числу всех разных производных существительных (N'_1); 2) числу всех разных суффиксальных существительных (Sl). Значение рассматриваемых индексов соответственно: I) $\frac{Sl_A}{N'_1} = 0,II$; 2) $\frac{Sl_A}{Sl} = 0,I4$.

IУ Индексы, выражающие отношение числа всех разных существительных, образованных суффиксальным способом от основы имени числительного (Sl_{Nu}) к: I) числу всех разных производных существительных (N'_1); 2) числу всех разных суффиксальных существительных (Sl). Указанные индексы соответственно имеют значения: I) $\frac{Sl_{Nu}}{N'_1} = 0,002I$; 2) $\frac{Sl_{Nu}}{Sl} = 0,0027$.

Рассматривая приведенные индексы, можно заметить, что если они связаны с производящей глагольной основой, то их значения значительно больше, чем у индексов, которые показывают отношение любой другой производящей основы к числу

существительных любого вида. Например, если сравнить индекс показывающий отношение числа существительных, образованных от производящей глагольной основы к числу всех суффиксальных существительных 0,55, то видно, что он почти в два раза больше, чем индекс $\frac{Sl_{N}}{Sl}$ и в 4 раза больше, чем $\frac{Sl_{A}}{Sl}$, значение которых соответственно 0,29 и 0,14.

Несколько индексов указывают на параметры, связанные со словообразовательной валентностью производящих основ суффиксальных существительных. Рассмотрим эти индексы.

I Индексы, выражающие отношение числа всех разных существительных, образованных суффиксальным способом от поливалентной производящей основы (PVal) к: 1) числу всех разных существительных (N'); 2) числу всех разных производных существительных (N'_1); 3) числу всех разных суффиксальных существительных (Sl). Значение рассматриваемых индексов соответственно: 1) $\frac{PVal}{N'}$ = 0,15; 2) $\frac{Pval}{N'_1}$ = 0,27; 3) $\frac{Pval}{Sl}$ = 0,34.

II Индексы, выражающие отношение числа всех разных существительных, образованных суффиксальным способом от моновалентной производящей основы (M_{val}) к: 1) числу всех разных существительных (N'); 2) числу всех разных производных существительных (N'_1); 3) числу всех разных суффиксальных существительных (Sl). Приведенные индексы имеют соответственно значения: 1) $\frac{Mval}{N'}$ = 0,27; 2) $\frac{Mval}{N'_1}$ = 0,51; 3) $\frac{Mval}{Sl}$ = 0,65.

Сравнивая эти индексы, видим, что индекс уменьшается, если увеличивается число существительных (N', N'_1 или Sl), соотносимых с числом существительных одной валентности. Например, у $\frac{Pval}{N'}$ индекс 0,15, так как N' = 1719. Напротив, у $\frac{Pval}{Sl}$ индекс 0,34, так как Sl = 757. Во всех случаях значения индексов моновалентных основ больше, чем у индексов поливалентных. Например, индекс $\frac{MVal}{Sl}$ = 0,65 на 1,8 раз больше, чем аналогичный индекс $\frac{PVal}{Sl}$.

Существенное значение имеют также параметры, выражаемые рядом индексов, в которых отражается различие в числе основ разных существительных, образованных суффиксальным способом в зависимости от количества морфем их составляющих. Рассмотрим эти индексы.

I Индексы, выражающие отношение числа всех разных суффиксальных существительных, у которых производящая основа состоит только из одной морфемы (M_1) к:
1) числу всех разных существительных (N'); 2) числу всех

разных производных существительных (N'_1); 3) числу всех разных суффиксальных существительных (S1). Приведенные индексы соответственно имеют значения: I) $\frac{M_1}{N'}$ = 0,I8; 2) $\frac{M_1}{N'_1}$ = = 0,33; 3) $\frac{M_1}{S1}$ = 0,42.

II Индексы, выражающие отношение числа всех разных существительных, образованных суффиксальным способом, у которых в состав производящей основы входят две морфемы (M_2) к: I) числу всех разных существительных (N'); 2) числу всех разных производных существительных (N'_1); 3) числу всех разных суффиксальных существительных (S1). Эти индексы соответственно имеют значения: I) $\frac{M_2}{N'}$ = 0,I0; 2) $\frac{M_2}{N'_1}$ = 0,I8; 3) $\frac{M_2}{S1}$ = 0,23.

III Индексы, выражающие отношение числа всех разных существительных, образованных суффиксальным способом, в состав производящей основы которых входят три морфемы (M_3) к: I) числу всех разных существительных (N'); 2) числу всех разных производных существительных (N'_1); 3) числу всех разных суффиксальных существительных (S1). Рассматриваемые индексы соответственно имеют значения: I) $\frac{M_3}{N'}$ = 0,03; 2) $\frac{M_3}{N'_1}$ = 0,07; 3) $\frac{M_3}{S1}$ = 0,08.

IУ Индексы, выражающие отношение числа всех разных существительных, образованных суффиксальным способом, в состав производящей основы которых входят четыре морфемы (M_4) к: I) числу всех разных существительных (N'); 2) числу всех разных производных существительных (N'_1); 3) числу всех разных суффиксальных существительных (S1). Указанные индексы имеют соответственно значения: I) $\frac{M_4}{N'}$ = 0,006; $\frac{M_4}{N'_1}$ = 0,0II; $\frac{M_4}{S1}$ = 0,0I4.

Приведенные индексы показывают, что более высокие значения имеем в тех случаях, когда индексы выражают отношение суффиксальных существительных с одноморфемной основой к числу существительных любого типа, или же отношение числа слов с любым количеством морфем в основе к числу всех суффиксальных существительных.

Другой параметр, характеризующий суффиксальные существительные, это концентрация их частоты. Она высчитывалась по формуле $K = \frac{\sum X_i}{F^*}$ где X_i абсолютная накопленная частота слов образованных при помощи I6 высокочастотных суффиксов, а F^* – абсолютная накопленная частота всего обследуемого массива текстов. Параметр К показан в таблице № 2, где приводятся К по каждому из пяти памятников, которые составляют выборочную совокупность.

Памятник	$\frac{X_f}{F_f}$	K
БД	$\frac{842}{120000}$	0,007016
СД	$\frac{857}{120000}$	0,007141
КС	$\frac{527}{120000}$	0,004391
КД	$\frac{724}{120000}$	0,006033
ПС	$\frac{847}{120000}$	0,007058

Из таблицы № 2 видим, что концентрация суффиксальных существительных в БД, СД и ПС одинакова (К = 0,007), весьма близка в КД (К = 0,006) и несколько меньше в КС (0,04). В целом такой существенный для сопоставления отдельных текстов параметр как К подтверждает большую близость обследуемых памятников с точки зрения насыщения текста суффиксальными существительными.

Таким образом, в статье были показаны такие индексы характеризующие словообразование существительных, как деривации, морфемности, валентности, концентрации. В то же время некоторые параметры, как индексы повторяемости, а также дисперсионные характеристики не рассматривались, так как они могут служить предметом отдельной статьи.

ЛИТЕРАТУРА

Бектаев К.Б., Лукьяненков К.Ф. О законах распределения единиц письменной речи. – Статистика речи и автоматический анализ текста. Л., 1971.

Каширина М.Е. О типах распределения лексических единиц в тексте. – Статистика речи и автоматический анализ текста. Л., 1974.

Надарейшвили И.Ш., Орлов Ю.К. Метод полной фиксации текста при лингвостатистическом анализе. – Проблемы общей и прикладной лингвистики. Тарту, 1978.

Нешитой В.В. Законы распределения слов в тексте и его лексическая параметризация. АКД, Минск, 1973.

Орлов Ю.К., Надарейшвили И.Ш. Рост лексики как функция длины текста. – Сообщения АН ГССР, т. 64, № 3, 1971.

Рускова М.П. Статистическое распределение лексики в болгарской письменности XVIII в. – Структурная и прикладная лингвистика. Л., 1978.

Рускова М.П. Частота знаменательных частей речи в болгарской письменности XVIII в. - Структурная и прикладная лингвистика, вып. 2. Л., 1983.
Тулдава Ю. О статистической структуре словаря. - Linguistica, III, Tartu, 1976.
Тулдава Ю. К вопросу об аналитическом выражении связи между объемом словаря и объемом текста. - Труды по лингвостатистике, VI, Тарту, 1980.
Якубайтис Т.А. Части речи и типы текстов. Рига, 1981.
Guiraud P. Les caractères statistiques du vocabulaire. Paris, 1954.
Petruszewycz M. Description statistique de textes littéraires russes par les méthodes de Markov. - Revue des études slaves, t. 55, fasc. 1, Paris, 1983.
Těšitelová M. Nouns in Lexical Statistics. - Prague Studies in Mathematical Linguistics. v. 2. Prague, 1966.
Těšitelová M. Využití statistických metod v grammatice. Praha, 1980.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

Беленски дамаскин. Ркп. № 713 (445). Народная библиотека в Софии, л. 10-99об.
Котленски дамаскин. Ркп. № 13.5.18. Библиотека АН СССР в Ленинграде.
Кованлъшки сборник. Ркп. № 13.6.16. Библиотека АН СССР в Ленинграде, л. 14-201.
Сборник от XVIII в., написан в Пазарджик. Ркп. № 435 (635). Народная библиотека в Софии, л. 1-188.
Свищовски дамаскин. - В кн.: Български старини, кн. УП. София, 1923, с. 75-130.

STATISTICAL PARAMETERS FOR WORD-FORMATION IN THE 18TH CENTURY BULGARIAN LANGUAGE

Mira P. Ruskova

S u m m a r y

The paper deals with the characteristic parameters for word-formation of nouns in the 18th Bulgarian manuscripts. These parameters are numerous and they includes derivation, suffixation, morphems, valency and concentration. The correlation analysis, i.e. the determination of the significance of each component in the noun word-formation, is of great importance when analysing the above mentioned parameters.

ЧАСТОТНАЯ СТРУКТУРА ТЕКСТА И ЗАКОН ЦИПФА

Ю.А. Тулдава

В статье определяется понятие частотной структуры текста и рассматривается одна из ее сторон – т.наз. ранговое распределение лексических единиц. В связи с этим обсуждается проблема практического использования и теоретического обоснования закона Ципфа. Иллюстративным материалом служат данные исследований по эстонскому языку (частично в сравнении с данными из других языков).

Частотная структура текста. Если абстрагироваться от конкретных лексических единиц[+], составляющих частотный словарь (ЧС), и рассматривать лишь частоты (F_i) и ранги (i) лексических единиц, то получается т.наз. (частотно-)р а н г о в о е распределение, или распределение "ранг – частота". Другой возможностью формального анализа ЧС является сопоставление частот F_i с численностью (количеством) единиц, имеющих данную частоту – $m(F_i)$, что дает (частотно-)с п е к т р а л ь н о е распределение, или частотный спектр (именуемый также "лексико-частотным" или "лексическим" спектром). Совместное рассмотрение рангового и спектрального распределений раскрывает нам ч а с т о т н у ю с т р у к т у р у текста, представляющую собой определенный аспект общей статистической организации текста (охватывающей всю проблематику анализа структуры и функционирования текста и соответствующего словаря в квантитативном освещении).

Определяемую таким образом частотную структуру текста можно представить наглядно в следующей компактной форме:

[+]Под лексической единицей мы понимаем е д и н и ц у у ч е т а на лексическом уровне анализа. Такой единицей является слово в разных своих проявлениях (лексема, словоформа) и в разных планах рассмотрения (в словаре, в тексте). На определенных этапах квантитативного анализа различаются такие понятия (единицы) как лексема, словоформа, словоупотребление, но в тех случаях, когда различение этих понятий несущественно, употребляется общее название с л о в о.

i	F_i	$m(F_i)$
1	F_1	$m(F_1)$
2	F_2	$m(F_2)$
...	...	...
$k \div n$	$F_{k \div n}$	$m(F_{k \div n})$
...	...	...

На материале ЧС лексем авторской речи современной эстонской художественной прозы (Kaasik U. и др., 1977) комплексное распределение рангов, частот и численности слов с данной частотой по указанной схеме приводится в табл. I (в конце статьи). Наряду с комплексом ранжированных величин исходными данными при рассмотрении частотной структуры текста являются также объем текста (N) и объем соответствующего словаря ($\angle$). В данном конкретном случае $N = \sum F_i$ = = 99 898 (словоупотреблений) и $\angle = i_{max}$ = I4 654 (лексем) (см. табл. I).

Ранговое и спектральное распределения могут иметь дифференциальную (некумулятивную) форму, или они могут быть представлены в интегральной (кумулятивной) форме, т.е. в виде накопленных частот или численностей.+ Интегральное распределение выражает покрываемость текста, которая позволяет судить о степени концентрации лексических единиц в определенных участках ЧС. Естественно, что частоты и численности могут быть представлены как в абсолютных, так и относительных величинах (в т.ч. в процентах). Для примера приводим данные о частотной структуре эстонского текста (в сокращенной форме) вместе с данными о покрываемости текста лексемами и словоформами (табл. 2 и 3).

Частотную структуру текста можно рассматривать комплексно как одно целое (см., например, Арапов М.В., Ефимова Е.Н., 1975) или в ней можно выделить отдельные аспекты для более подробного анализа. В рамках данной статьи мы рассмотрим лишь одну из сторон частотной структуры - ранговое распределение и возможности его аналитического описания с помощью разных вариантов т.наз. закона Ципфа.

+ Соответственно различаются дифференциальная функция (плотность) и интегральная функция распределения (см. Митропольский А.К., I97I, с. 209 и след.). В ином аспекте рассматриваются дифференциальное уравнение (включающее производные или дифференциалы) и интегрирование (т.е. решение дифференциального уравнения).

Ранговое распределение и закон Ципфа. Одной из важнейших закономерностей, выявленных при квантитативном анализе текстов, является статистическая связь между ч а с т о т о й и р а н г о м слова. При этом констатируется, что, хотя в различных текстах слова могут иметь различные ранги, все же устойчивой является сама форма распределения, т.е. вид закономерности в целом. Здесь сказывается "топологический" принцип, согласно которому "важна не "метрика" /.../, но зато важно сохранение "схемы", которая может видоизменяться, оставаясь самой собой и может "наполняться" разным содержанием" (Бернштейн Н.А., 1966, с. 65). Во всех случаях, когда мы имеем дело с текстами естественного языка, проявляется т.наз. эффект к о н ц е н т р а ц и и и р а с с е я н и я, который состоит в том, что имеется небольшая группа очень частых слов ("ядро" ЧС) и большая группа редких слов ("хвост" ЧС); между ними наблюдается плавный переход ("зона среднечастотных слов"). На графике это выражается формой, напоминающей гиперболу (см., например, рис. I). Такая неравномерность в распределении единиц обнаруживается не только в отношении слов, но и в отношении других языковых единиц (букв, фонем, морфем, словосочетаний и т.д.). Более того, сходная форма распределения встречается и во многих других областях человеческой деятельности (информатике, экологии, демографии и др.), что заставляет рассматривать этот тип распределения как универсальный семиологический "закон предпочтения" (Перебийніс В.С., 1970), или "закон распределения единиц по значимости" (Мартыненко Г.Я., 1978).

Для аналитического выражения зависимости между частотой и рангом слова предлагается множество формул, которые представляют собой разновидности закона Ципфа. (Zipf G.K., 1935; 1949; Mandelbrot B., 1954; Орлов Ю.К., 1970; Алексеев П.М., 1978; Крылов Ю.К., 1982; и др.). О с н о в н а я ф о р м а распределения Ципфа выражается следующей формулой, которая представляет собой степенную функцию с отрицательным показателем степени:

$$F_i = C i^{-\gamma} \quad \text{или} \quad p_i = \kappa i^{-\gamma}, \qquad (I)$$

где F_i - абсолютная частота, p_i - относительная частота (вероятность), i - ранг, C, κ и γ - параметры распределения (причем $p_i = F_i/N$ и $C = \kappa N$, где N - объем текста). В частном случае, когда $\gamma = 1$, формула принимает вид ("классическое однопараметрическое распределение Ципфа"):

$$F_i = Ci^{-1} \quad \text{или} \quad p_i = \kappa i^{-1}. \qquad \text{(Ia)}$$

Каков же содержательный смысл основной формы закона Ципфа? Чтобы раскрыть "механизм" изменения переменных, рассмотрим соответствующее функции (I) дифференциальное уравнение, имеющее следующий вид:

$$\frac{dF_i/F_i}{di/i} = \lim \frac{\Delta F_i/F_i}{\Delta i/i} = \gamma \quad (\gamma < 0).$$

Из уравнения явствует, что отношение скоростей относительного изменения F_i и i остается постоянным.+

Так характеризуется и закон "постоянного относительного роста (или убывания)", известный во многих областях науки (см., например, Ланд К.Ч., 1977, с. 388). Таким образом, закон Ципфа совпадает по форме с неким универсальным законом, охватывающим широкий круг явлений материального мира. В данном конкретном случае констатируется, по существу, наличие такой связи между текстом и соответствующим словарем, упорядоченными особым образом, т.е. связи между рядом убывающих частот в тексте (F_i) и ростом словаря ($i = 1, 2, \ldots, L$).

Ранговое распределение, соответствующее функции (I), имеет на графике форму гиперболы, а в билогарифмических координатах оно должно давать прямую линию, т.е. линейную зависимость между $\ln F_i$ и $\ln i$. Это подтверждается, в частности, на материале ЧС словоформ эстонского языка (см. рис. 2). Известно, что типологическое различие между языками обнаруживается особенно ярко при сопоставлении ранговых распределений словоформ. Сравнивая, например, данные ЧС словоформ эстонского и английского языков (при равных объемах текстов), мы можем констатировать, что значения параметра γ существенно различаются: для флективно-синтетического эстонского языка $\gamma = 0{,}86$, а для флективно-аналитического английского языка $\gamma = 1{,}02$ (Kučera H., Francis W.N., 1967, с. 357), т.е. угол наклона прямой на графике (и, соответственно, тангенс угла γ) для английского языка значительно больше, чем для эстонского языка. Это означает, что высокочастотные словоформы (в основном служебные слова) в английском языке покрывают бóльшую часть текста, чем в эстонском языке, и запас словоформ заканчивается в английском тексте быстрее. В результате таких

+ Это уравнение можно переписать в виде $dF_i/F_i = -\gamma(di/i)$. Интегрируя его, получим $\ln F_i = c - \gamma \ln i$, где c – константа. Положив $c = \ln C$, получим $F_i = Ci^{-\gamma}$, т.е. функцию (1).

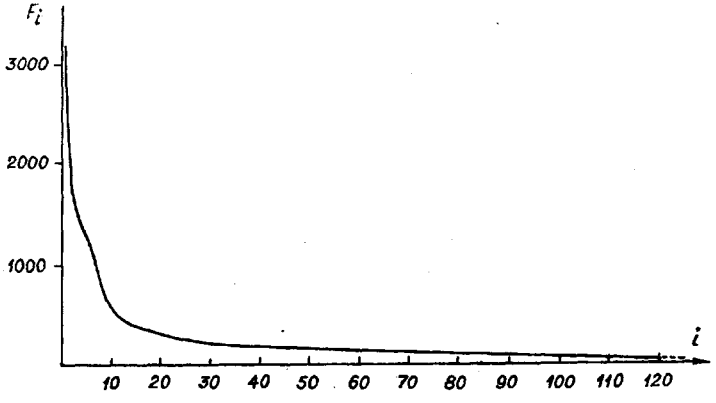

Рис. I. Связь между частотой (F_i) и рангом (i) слова по данным ЧС словоформ эстонского языка.

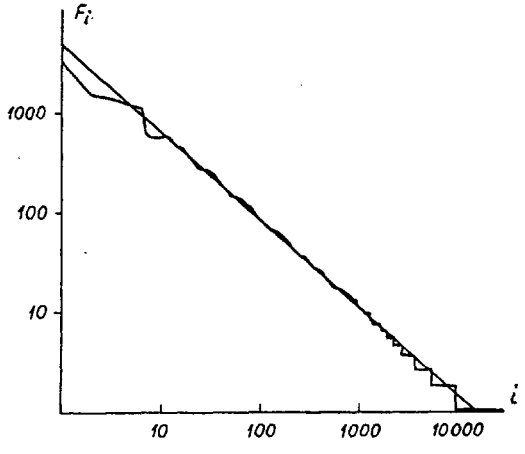

Рис. 2. Связь между F_i и i по данным ЧС словоформ. Билогарифмический масштаб.

лингвистико-типологических расхождений в эстонском тексте объемом 100 тыс. словоупотреблений имеется около 30 тыс. словоформ, в то время как в английском тексте такого же объема их число не превышает 15 тыс.

Отклонения от закона Ципфа. Поправка Мандельброта. Более подробный анализ показывает, что точная линейная зависимость между частотой и рангом (в билогарифмических координатах) выполняется не на всем протяжении рангового распределения. Как и во многих других языках, в эстонском языке обнаруживается отклонение от линейной зависимости в ранговом распределении языковых единиц в области б о л ь ш и х ч а с т о т (в зоне "ядра" ЧС). Для улучшения соответствия между эмпирическими и теоретическими данными обычно используется вариант закона Ципфа, включающий т.наз. поправку Мандельброта (Mandelbrot B., 1954). Этот вариант называется "каноническим законом" Ципфа или "законом Ципфа-Мандельброта" (название канонического закона дал ему Б. Мандельброт; см. Плат У., 1965, с. 206):

$$F_i = C(i+B)^{-\gamma} \quad \text{или} \quad p_i = \kappa(i+B)^{-\gamma}, \qquad (2)$$

где B - поправка Мандельброта. В частном случае, когда $\gamma = 1$, формула имеет вид

$$F_i = C(i+B)^{-1} \quad \text{или} \quad p_i = \kappa(i+B)^{-1}. \qquad (2а)$$

По данным эстонского языка значение поправки Мандельброта B для ЧС лексем равняется 2,3, а для ЧС словоформ $B = 0{,}5$, т.е. отклонение в зоне ядра ЧС словоформ несколько меньше, чем в соответствующей зоне ЧС лексем. Формула (2) дает хорошее соответствие между теоретическими и эмпирическими данными в начальной части ЧС, а при $i > 15$ практически не меняет общий вид зависимости, причем применение поправки Мандельброта приводит к некоторому увеличению теоретических значений констант C (или κ) и γ (см. табл. 4).

Кроме отклонения от линейной зависимости (в билогарифмических координатах) в головной части ЧС, во многих случаях, в частности и в отношении ЧС лексем эстонского языка, наблюдается более или менее заметное отклонение от линейной зависимости в области м а л ы х ч а с т о т (в "хвосте" распределения). По данным сводного словаря лексем такое отклонение небольшое (при $i = 30 \div 2500$ тангенс угла наклона $\gamma = 0{,}98$, а при $i > 2500$ $\gamma = 1{,}03$; см. табл. 5). Откло-

нение в области малых частот более заметное по данным сводного словаря лексем русского языка в тексте такого же объема (табл. 5). Оно заметно и в индивидуальном словаре эстонского языка (табл. 6).

Отклонения от линейной зависимости (в билогарифмических координатах) в начальной и конечной частях рангового распределения, как правило, увеличиваются с нарастанием объема выборки (текста). Некоторые исследователи считают, что закон Ципфа может выполняться лишь на единственном значении объема однородной выборки, причем этот объем связывается с понятием "целостности" и даже "литературной завершенности" текста (см., например, Орлов Ю.К., 1978). Во всяком случае, можно утверждать, что частотная структура текста динамична и закономерно изменяется с увеличением объема текста. Например, по данным эстонского языка, отклонения от основного варианта закона Ципфа (по формуле (I)) в малых выборках противоположны отклонениям в больших выборках (Тулдава Ю., 1977, с. 162; см. также Орлов Ю.К., 1978). Такая "динамика" частотной структуры текста наблюдается как в отношении сводных, так и индивидуальных текстов. Таким образом, отклонения от закона Ципфа в его основной форме (I) в некоторых (гл. обр. больших) текстах можно считать вполне закономерным явлением, учитывая динамический характер статистической организации текста. Представляется, что наиболее общим условием выполнения закона Ципфа в своей ранговой форме следует считать линейность связи между $\ln F_i$ и $\ln i$ в средней части распределения, в то время как отклонения в начальной и конечной частях, а также конкретное значение параметра γ объясняются разными, в т.ч. лингвистическими причинами (тип языка, структурные особенности словаря, выбор единиц и др.; см., например, Борода М.Г., Поликарпов А.А., 1984).

Особой точки зрения придерживается Г.Я. Мартыненко (1978), который считает, что упомянутое распределение, как распределение неоднородных элементов, в принципе не может быть аппроксимировано единой функцией; в частности ядерные и периферийные элементы описываются разными законами. Однако другие исследователи продолжают разработку новых обобщающих моделей ципфовского распределения.

Тезис о нелинейности связи. В связи с констатацией отклонений от линейной зависимости переменных (в билогарифмическом масштабе) представляет интерес тезис П.М. Алексеева (1978) о "криволинейном" характере рангового распределения

лексических единиц в больших текстах. Предполагается, что при большом увеличении данного текста многие редкие слова передвигаются в средние ярусы частотного словаря и происходит изменение характера зависимости ранг - частота: платформы в нижних ярусах уменьшаются, зато они будут расширяться в средней части словаря, сдвигая эту часть теоретического графика вправо и увеличивая его кривизну (см. также: Пиотровский Р.Г., 1975, с. 107-110). П.М. Алексеев предлагает следующую обобщающую формулу закона Ципфа:

$$F_i = C i^{-(\gamma+\varphi \ln i)} \quad \text{или} \quad p_i = \kappa i^{-(\gamma+\varphi \ln i)}, \qquad (3)$$

где C (или κ), γ и φ - параметры. Эта формула в логарифмической записи соответствует уравнению параболической регрессии. П.М. Алексеев называет выражение (3) "четвертым приближением" закона Ципфа, имея в виду, что первым приближением является классическая однопараметрическая форма (формула (Ia)), а вторым и третьим приближениями - формулы (I) и (2).

Формула (3) обладает тем преимуществом, что она включает в себя основные формы распределения Ципфа в качестве частных случаев: при $\varphi = 0$ формула (3) превращается в формулу (I), а если при этом $\gamma = 1$, то получается однопараметрическая формула (Ia). Проверка показывает, что формула (3) хорошо аппроксимирует эмпирические данные в том случае, если имеются отклонения от "линейной" зависимости в начальной и конечной частях распределения. Вопрос только в том, отражает ли формула (3) д е й с т в и т е л ь н у ю структуру текста (параболическая зависимость между F_i и i в билогарифмических координатах), или она лишь "затушевывает" факт закономерных отклонений в начальной и конечной частях распределения, игнорируя то, что в средней части распределения в действительности сохраняется линейность связи (на графике прямая линия в билогарифмических координатах). Для окончательного решения проблемы потребуются новые исследования на материале больших массивов текстов. Пока что представляется, что линейность связи в средней части распределения в больших текстах все же сохраняется, если судить, например, по графикам распределения словоформ и лексем в некоторых английских и французских текстах объемом от I до 71 млн. словоупотреблений (Алексеев П.М., 1983, с. 50, рис. 14, и с. 55, рис. 18).

Дальнейшее теоретическое осмысление "нелинейной" концепции закона Ципфа дано в работе В.Н. Бичкова (1984), который развивает тезис о "расщеплении" параметра γ на две составляющие - собственно константу γ_0 и переменную, "скользящую" γ_i. Формулу (3) можно переписать в виде

$$F_i = C i^{-\gamma_i}, \qquad (3а)$$

где γ_i (т.е. значение γ-параметра в конкретной точке частотно-рангового ряда) определяется как

$$\gamma_i = \gamma_0 e^{\alpha i},$$

где α - коэффициент прироста γ-параметра за интервал перехода от $i = 1$ до $i_{max} = V$.

"Двусторонние" варианты закона Ципфа. Основываясь на некоторых вероятностно-комбинаторных рассуждениях, Ю.К. Крылов (1982) выводит формулу связи между частотой и рангом слова (в наших обозначениях)

$$F_i = \frac{C}{i + B_1} - B_2, \qquad (4)$$

где C, B_1 и B_2 - параметры. Легко видеть, что эта формула отличается от формулы Ципфа-Мандельброта при $\gamma = 1$ (2а) лишь наличием слагаемого B_2. Параметры B_1 и B_2 обеспечивают возможность сдвига в направлениях, параллельных координатным осям, благодаря чему учитываются отклонения от "линейной" связи как в начальной, так и в конечной части распределения. Подразумевается, что ципфовский параметр γ, указывающий на уклон прямой в билогарифмических координатах, равняется единице (ср. формулу 2а)). В тех случаях, когда параметр γ по эмпирическим данным близок к единице (например, при $\gamma = 0{,}92$ для ЧС лексем эстонского языка), формула достаточно хорошо описывает связь между F_i и i.

Факт отклонения от линейной зависимости (в билогарифмических координатах) рангового распределения послужил поводом для выведения особого варианта закона Ципфа также польским исследователем Е. Ворончаком (Woronczak J., 1967). В его формуле учитываются отклонения как в начальной, так и в конечной частях распределения, так же как и у Ю.К. Крылова (см. формулу (4)), но в отличие от последнего Е. Ворончак предлагает вариант, в котором параметр γ (коэффициент убывания) может отличаться от единицы ($\gamma \neq 1$). В наших обозначениях формула Е. Ворончака имеет вид:

$$F_i = N(i+B)^{-\gamma} z^i \phi^{-1}, \qquad (5)$$

где γ, B и z – параметры, N – объем текста,

$$\phi = \sum_{i=0}^{\infty} (i+B)^{-\gamma} z^i .$$

Параметр B играет роль поправки Мандельброта, учитывающей отклонение в начальной части распределения, а параметр z (вернее, выражение z^i при $|z| < 1$) обеспечивает более быстрое убывание функции по сравнению с формулой Мандельброта (2); тем самым учитывается отклонение в конечной части распределения. По существу, множитель z^i имеет такое же действие, как и возрастание γ с рангом i. Этот принцип в некоторой степени роднит метод Е. Ворончака с "нелинейным" подходом П.М. Алексеева и В.Н. Бычкова (см. формулы (3) и (3а)), и подобно им, Е. Ворончак по существу игнорирует "линейный" характер связи на среднем участке распределения.

Концепция "объема Ципфа". Формула Мандельброта в своей двухпараметрической форме (2а) является основой для "обобщенного закона Ципфа-Мандельброта" Ю.К. Орлова (1970; 1976; 1978). Этот автор развивает оригинальный подход, при котором центральным понятием является т.наз. "объем Ципфа" (обозначаемый символом Z), служащий отправной точкой для вычисления параметров частотной структуры текста. "Объем Ципфа" указывает по замыслу Ю.К. Орлова на тот единственный объем данного текста, при котором теоретически может выполняться закон Ципфа-Мандельброта. Если взять за основу фактический объем текста (т.е. если считать, что $Z = N$), то формула для вычисления частот слов принимает вид (в наших обозначениях):

$$F_i\,(Z = N) = C_1 (i + B_1)^{-1}, \qquad (6)$$

где $C_1 = N(\ln F_1)^{-1}$; $B_1 = CF_1^{-1} - 1$; F_1 – частота наиболее частотного слова в данной выборке. Если же исходить из теоретически вычисляемого значения Z (о вычислениях см. Орлов Ю.К., 1978), то формула принимает вид

$$F_i\,(Z \neq N) = C_2 (i + B_2)^{-1}, \qquad (6а)$$

где $C_2 = N\left[\ln(Z p_1)\right]^{-1}$ и $B_2 = CF_1^{-1} - 1$; p_1 – относительная частота наиболее частотного слова.

Отметим, что параметр γ, указывающий на наклон прямой в билогарифмических координатах, в этих формулах жестко определен и равен единице ($\gamma = 1$). (О формулах при $\gamma \neq 1$ см. Орлов Ю.К., 1976, с. 184 и след.).

Формулу Мандельброта (2а), т.е. вариант с $\gamma = 1$, а также формулы Орлова (6) и (6а), можно переписать в следующем виде:

$$F_i^{-1} = (i + B)C^{-1} = iC^{-1} + BC^{-1}.$$

Положив $BC^{-1} = a$ и $C^{-1} = b$, получаем

$$F_i^{-1} = a + bi, \qquad (7)$$

т.е. констатируется наличие линейной связи между рангом i и обратной величиной F_i. В явном виде

$$F_i = (a + bi)^{-1}. \qquad (7а)$$

На материале ЧС лексем эстонского языка линейная связь между F_i^{-1} и i действительно имеет место (см. рис. 3), благодаря тому, что в данном случае значение параметра γ близко к единице. Соответствие между наблюдаемыми и ожидаемыми данными хорошее (см. табл. 7). Формулы (7) и (7а) представляют собой не только равноправные альтернативы к формулам типа (2а), но они позволяют также более простым способом вычислить параметры C и B (нет необходимости использования итерационного метода). Кроме того, вычисление параметров по формуле (7) методом наименьших квадратов (или графическим способом) является удобным и наглядным способом проверки точности формул типа (2а), особенно в тех случаях, когда значения параметров найдены теоретическим путем. Например, на графике (рис. 3) видно, что выравнивающая прямая (II), вычисленная теоретически на основе "объема Ципфа", не совсем точно следует эмпирическим точкам, более того, здесь отклонение носит явно систематический характер.[+] В данном случае

[+] Вычисления на основе "объема Ципфа" по формуле (6а) дают (при $p_1 = 0{,}0424$ и $Z = 175\,000$): $C = 11200$ и $B = 1{,}65$; следовательно, $a = BC^{-1} = 1{,}47 \cdot 10^{-4}$ и $b = C^{-1} = 0{,}89 \cdot 10^{-4}$ (см. прямую II на рис. 3). В действительности же, по формуле (7): $a = 1{,}5 \cdot 10^{-4}$ и $b = 0{,}75 \cdot 10^{-4}$ и, соответственно, $C = b^{-1} = 13300$ и $B = aC = 2$.

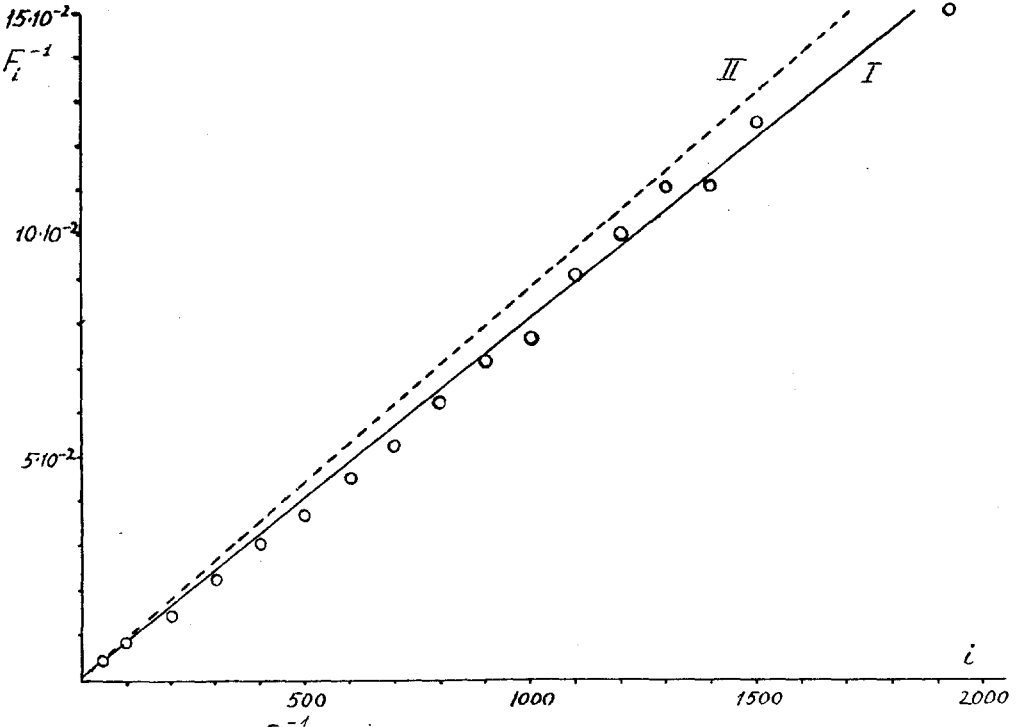

Рис. 3. Связь между F_i^{-1} и i по данным ЧС лексем эстонского языка. Выравнивание по формуле (7): $F_i^{-1} = \alpha + \beta i$ (I) и по формуле (6а) на основе "объема Ципфа" (II).

мы имеем дело со сводным текстом одного подъязыка, а вычисления по "объему Ципфа" Ю.К. Орлов рекомендует произвести на основе индивидуальных текстов, считая, что индивидуальные тексты обладают какими-то особыми свойствами по сравнению со сводными текстами. Однако наш опыт показывает, что сама техника вычисления "объема Ципфа" страдает существенными недостатками (см. Тулдава Ю., 1980, с. 115 и след.). Нельзя быть также вполне уверенным в правомерности исходных постулатов об особом месте и значении "объема Ципфа" в статистической организации текста.

Дискретный аналог закона Ципфа. Все рассмотренные выше варианты формул закона Ципфа (в ранговой форме) представляют собой непрерывные функции, хотя в действительности мы имеем дело с дискретным распределением лингвистических объектов. Обычно игнорируется этот факт, так как считается, что непрерывная функция вполне адекватно отражает все особенности "скачкообразно" изменяющейся эмпирической функции частотно-рангового распределения. Желая все же уточнить вид распределения, М.В. Арапов и др. (1975) предлагают особый "дискретный аналог закона Ципфа", при котором учитываются дополнительное условие - целочисленность частоты или ранга. Соответствующая формула имеет вид:

$$F_i = \frac{\beta(L+1)^{\gamma}}{1-\gamma}\left[(i+1)^{1-\gamma} - i^{1-\gamma}\right], \quad (8)$$

где параметр β определяется в зависимости от значения ципфовского параметра γ (см. Арапов М.В. и др., 1975, с. 6), L - объем словаря (у авторов обозначение N). На основе ЧС лексем эстонского языка (при $\gamma = 0{,}92$; $\beta = 0{,}6562$; $L = 14654$) формула (8) приближенно описывает данное эмпирическое частотно-ранговое распределение, но лучшее соответствие достигается при анализе ЧС словоформ (см. табл. 8).

Здесь, как и в других случаях теоретического вывода распределений, обнаруженные отклонения от эмпирических данных объясняются естественной флуктуацией значений параметров лингвистических распределений, которые мы рассматриваем как вероятностные системы с моментами устойчивости и вариативности. Но во всех случаях сохраняется общий вид распределения, характеризующийся концентрацией и рассеянием объектов.

Обоснование закона Ципфа. Широкая применимость закона Ципфа (в разных своих вариантах) к описанию подобных распределений в лингвистике и в других областях социальной жизни

человека привела к появлению множества работ по теоретическому обоснованию этого закона. Известно, что сам Ципф (Zipf G.K., 1935; 1949) выдвинул гипотезу "наименьших усилий" для объяснения единообразного характера частотных распределений событий в процессах передачи информации. Многие другие авторы исходят также из принципа оптимальности в функционировании языка, как самоорганизующейся системы, учитывая, что "язык формировался на протяжении тысячелетий в результате коллективного творчества миллиардов людей" (Козачков Л.С., 1969, с. 11).

Доказательством того, что распределение лингвистических единиц по закону Ципфа является результатом процесса эволюционного развития, приводятся, например, факты отклонения от этого закона, наблюдаемые в речевой деятельности детей. По данным специальных исследований (Marx W., Schürer-Necker E., 1978), кривая рангового распределения слов для речи маленьких детей более пологая (параметр $\gamma \ll 1$ в билогарифмическом масштабе), чем для речи старших детей и взрослых. Из этого делается вывод, что в речи маленьких детей энтропия более высокая и что развитие речи, по-видимому, идет по пути уменьшения энтропии (и тем самым, по пути увеличения избыточности, что связывается в данном случае с оптимизацией декодирования). Это согласуется со стремлением самоорганизующихся открытых систем к варианту, характеризующемуся минимальной энтропией, хотя, в действительности, в зависимости от конкретных условий, энтропия стабилизируется на каком-то оптимальном для данных условий уровне (в человеческой речи при $\gamma \approx 1$).

Считается также, что выполнение закона Ципфа свидетельствует о высокой степени организованности целостного текста (Орлов Ю.К., 1970; 1976; 1978), о его "правильности" (Арапов М.В. и др., 1975). Делается попытка связать процесс порождения текста с идеями теории классификации, формулируется принцип "минимума симметрии" естественной классификационной системы и показывается, что из этого принципа может быть выведен закон Ципфа (Арапов М.В., Шрейдер Ю.А., 1978).

Кроме признания "оптимизирующего" начала при выводе закона Ципфа, этот закон пытаются вывести, используя понятие вероятностного процесса (например, Simon H.A., 1955; Hill B.M., 1982) или прибегая к аналогии с термодинамикой (Mandelbrot B., 1954; Шрейдер Ю.А., 1967; Крылов Ю.К., 1982). Из новейших попыток можно отметить оригинальный подход Ю.К.

Крылова и М.Д. Якубовской (Krylov Yu.K., Yakubovskaya M.D., 1983), которые приходят к своеобразному обоснованию закона Ципфа на основе тезиса о порождении связного текста как периодического процесса, представляющего собой совокупность "волн" вероятностной природы, квадраты амплитуды которых определяют вероятности обнаружения того или иного фиксированного элемента в фиксированной точке процесса текстообразования.

В последнее время, гл. обр. благодаря исследованиям А.И. Яблонского (1976) и С.Д. Хайтуна (1983), утвердилось мнение, что закон Ципфа играет в соответствующих областях (в частности при описании социальных явлений) ту же универсальную роль предельного распределения, что и гауссов закон (нормальное распределение) в неорганических и др. процессах. В этой связи говорят о "негауссовости" закона Ципфа, содержательно выражающейся в явлении концентрации и рассеянии, а формально в отсутствии дисперсии (она равна бесконечности). Утверждается, что распределение Ципфа является не только одним из многих эмпирических распределений, а т е о р е т и ч е с к и м з а к о н о м, имеющим надежную математическую базу в виде теории устойчивых "негауссовых" распределений (Яблонский А.И., 1976). Известен также тезис о том, что распределение Ципфа представляет собой универсальный закон, сферой действия которого являются "естественно возникшие сложные системы" (Арапов М.В., Шрейдер Ю.А., 1978, с. 75). Постоянство (устойчивость) распределения Ципфа применительно к социальным явлениям дает основание предположить, что явления, подчиняющиеся закону Ципфа, можно рассматривать как системы, находящиеся в равновесном, т.е. в наиболее благоприятном для системы состоянии. Учитывая динамичность распределения и закономерную флуктуацию частот, ципфовскую структуру текста можно в целом охарактеризовать как п о д в и ж н о е р а в н о в е с и е системы.

Таким образом, и через системно-вероятностные рассуждения (негауссовость, подвижное равновесие) можно прийти к выводу об оптимизирующем характере закона Ципфа. В то же время, как нам представляется, надо остерегаться чрезмерного увлечения идеей оптимальности закона Ципфа, например, когда выполнение этого закона на языковом материале связывают с эстетическими свойствами текста, "художественной полноценностью" и т.п. Надо учесть то, что с помощью закона Ципфа "измеряется" только абстрактная, формально-частотная струк-

тура (упорядоченность элементов строении) текста, вне связи его с конкретным языковым заполнением. Таким образом, зависимость между частотной структурой и содержательным аспектом текста может быть только очень косвенной, опосредствованной, над установлением которой лингвистам предстоит еще очень много поработать.

Особо заслуживают внимания попытки связать принцип ципфовского распределения потоков речи ("концентрация и рассеяние") с деятельностью мозга. Например, А.Н. Лебедев (1983) пытается объяснить квантитативные особенности порождения речи пространственно-временной организацией периодических процессов головного мозга. Исходя из предположения о кодировании образов слов "пакетами волн нейронной активности", он сначала по аналогии выводит формулу для определения "полного диапазона колебаний ранга слова" (q) и затем приходит к выражению (в наших обозначениях):

$$F = CQ, \qquad (9)$$

где $C = F_{max}$ и $Q = \frac{1}{q} \ln(1+\frac{q}{i})$; i – среднее значение ранга слова. Тем самым построена формула, параметры которой по замыслу ее автора имеют ясный психологический и физиологический смысл. Близость опытных и теоретических данных подтверждается в предварительном эксперименте (Лебедев А.Н., 1983, с. 16), хотя вычисления кажутся несколько громоздкими. Но в данном случае для нас важна констатация того факта, что предположения о связи между особенностями частотной структуры текста и некоторыми закономерностями деятельности мозга, по-видимому, не лишены основания, и исследования в этой области приобретают актуальность.

Заключение. Подытоживая сказанное, следует подчеркнуть важность закона Ципфа при описании и объяснении некоторых существенных фактов, относящихся к статистической организации текста. В зависимости от конкретных условий (тип языка, выбор единиц, объем текста и др.) ранговое распределение лексических единиц может быть описано основным вариантом закона Ципфа, т.е. обыкновенной степенной функцией с отрицательным показателем степени (формула (I)), или одной из разновидностей (модификаций) закона Ципфа, рассмотренных в данной статье. Наиболее общим условием выполнения закона Ципфа можно считать "линейность" (в билогарифмических координатах) в средней части распределения лексических единиц, причем возможные отклонения в начальной и конечной частях объясня-

ются динамическим характером статистической организации текста и естественным образом входят в общую модель частотной структуры текста. В то же время нельзя отрицать и теоретической возможности нелинейной (криволинейной) связи в текстах очень большого объема (по П.М. Алексееву). По существу, закон Ципфа отражает в специфической форме общесемиологический "закон предпочтения" (В.И. Перебейнос), приводящий к феномену концентрации и рассеяния единиц. В функциональном смысле закон Ципфа (в своей основной форме) соответствует принципу "постоянного относительного роста (или убывания)" сравниваемых величин.

По всей вероятности, закон Ципфа, применительно к частотной структуре текста и, тем самым, к порождению речи, отражает результат естественного материального процесса эволюционного развития человеческого языка. Есть основание предполагать, что закон Ципфа согласуется со стремлением определенного типа самоорганизующихся систем к варианту, характеризующемуся оптимальным уровнем энтропии и "подвижным равновесием". Можно также предположить, что принципы функционирования, которые в общей форме отражает закон Ципфа, связаны с деятельностью мозга, его структурой. А деятельность мозга, в свою очередь, не может не отражать некоторых более общих закономерностей, присущих материальному миру в целом. Тем самым можно, видимо, говорить об универсальности закона Ципфа, применимости его к самым разнообразным отраслям реальной действительности.

Таблица I

Частотная структура текста по данным 40 авторов эстонской художественной прозы (i - ранг, P_i - частота, $m(P_i)$ - количество слов с данной частотой). Объем текста N - 99 898; объем словаря L - I4 654.

i	P_i	$m(P_i)$	i	P_i	$m(P_i)$	i	P_i	$m(P_i)$	i	P_i	$m(P_i)$	i	P_i	$m(P_i)$	i	P_i	$m(P_i)$
I	4837	I	4I	264	I	80	I53	I	I32	96	I	239-24I	57	3	587-6I2	22	26
2	3493	I	42	26I	I	8I	I5I	I	I33-I35	94	3	242-244	56	3	6I3-650	2I	38
3	2598	I	43	260	I	82	I50	I	I36-I39	93	4	245-247	55	3	65I-676	20	26
4	I98I	I	44	254	I	83	I49	I	I40-I4I	92	2	248-252	54	5	677-707	I9	3I
5	I395	I	45	248	I	84	I48	I	I42-I44	9I	3	253-256	53	4	708-747	I8	40
6	I300	I	46	238	I	85	I46	I	I45-I47	90	3	257-263	52	7	748-793	I7	46
7	I047	I	47	237	I	86	I44	I	I48-I49	89	2	264-268	5I	5	794-856	I6	63
8	879	I	48	234	I	87	I39	I	I50	88	I	269-270	50	2	857-893	I5	37
9	845	I	49	230	I	88	I38	I	I5I	87	I	27I-274	49	4	894-940	I4	47
I0	827	I	50	223	I	89	I36	I	I52-I54	86	3	275-280	48	6	94I-I0I7	I3	77
II	784	I	5I	2I8	I	90	I35	I	I55-I57	85	3	28I-285	47	5	I0I8-I080	I2	63
I2	634	I	52	209	I	9I-92	I33	2	I58-I59	84	2	286-289	46	4	I08I-II69	II	89
I3	6I3	I	53	207	I	93-94	I3I	2	I60	83	I	290-299	45	I0	II70-I293	I0	I24
I4	58I	I	54	206	I	95-96	I30	2	I6I-I62	82	2	300-305	44	6	I294-I4I4	9	I2I
I5	568	I	55	200	I	97	I26	I	I63-I64	8I	2	306-3I2	43	7	I4I5-I567	8	I53
I6	499	I	56	I98	I	98-99	I25	2	I65-I69	79	5	3I3-3I8	42	6	I568-I779	7	2I2
I7	496	I	57	I95	I	I00	I24	I	I70-I72	78	3	3I9-329	4I	II	I780-2044	6	265
I8	493	I	58	I94	I	I0I	I23	I	I73-I79	76	7	330-339	40	I0	2045-2389	5	345
I9	465	I	59	I92	I	I02-I03	I22	2	I80-I83	75	4	340-346	39	7	2390-2980	4	59I
20	448	I	60	I90	I	I04	I20	I	I84-I87	74	4	347-349	38	3	298I-39I8	3	938
2I	436	I	6I	I89	I	I05	II6	I	I88-I9I	73	4	350-360	37	II	39I9-5972	2	2054
22	434	I	62-63	I85	2	I06	II3	I	I92-I94	72	3	36I-369	36	9	5973-I4654	I	8682
23	428	I	64	I8I	I	I07-I08	II2	2	I95-202	7I	8	370-382	35	I3			
24	382	I	65	I80	I	I09-II0	II0	2	203-204	70	2	383-389	34	7			
25	373	I	66	I79	I	III	I09	I	205-206	69	2	390-404	33	I5			
26	366	I	67	I76	I	II2	I08	I	207	68	I	405-4I2	32	8			
27	350	I	68	I75	I	II3-II5	I07	3	208-209	67	2	4I3-428	3I	I6			
28	339	I	69-70	I74	2	II6	I06	I	2I0-2II	66	2	429-445	30	I7			
29	327	I	7I	I7I	I	II7-II9	I05	3	2I2-2I7	64	6	446-466	29	2I			
30	309	I	72	I67	I	I20	I04	I	2I8-220	63	3	467-479	28	I3			
3I-33	304	3	73	I66	I	I2I-I24	I0I	4	22I	62	I	480-509	27	30			
34	297	I	74	I65	I	I25-I26	I00	2	222-223	6I	2	5I0-529	26	20			
35-37	285	3	75-76	I64	2	I27	99	I	224-229	60	6	530-543	25	I4			
38-39	272	2	77-78	I63	2	I28	98	I	230-235	59	6	544-564	24	2I			
40	268	I	79	I59	I	I29-I3I	97	3	236-238	58	3	565-586	23	22			

Таблица 2

Частотно-ранговое распределение лексики по данным ЧС лексем авторской речи эстонской художественной прозы (i - ранг, P_i - частота, P_i^* - накопленная частота, P_i^* (%) - покрываемость текста)

i	P_i	P_i^*	P_i^*(%)	i	P_i	P_i^*	P_i^*(%)
I	4237	4237	4,24	400	33	59583	59,64
2	3493	7730	7,74	500	27	62517	62,58
3	2598	I0328	I0,34	600	22	64948	65,0I
4	I98I	I2309	I2,32	700	I9	66986	67,05
5	I395	I3704	I3,72	800	I6	68733	68,80
6	I300	I5004	I5,02	900	I4	70282	70,35
7	I047	I605I	I6,07	I000	I3	7I622	7I,70
8	879	I6930	I6,95	I00I - I0I7	I3	7I843	7I,92
9	845	I7775	I7,79	I0I8 - I080	I2	72599	72,67
I0	827	I8602	I8,62	I08I - II69	II	73578	73,65
20	448	24I23	24,I5	II70 - I293	I0	74818	74,89
30	309	27867	27,90	I294 - I4I4	9	75907	75,98
40	268	30743	30,77	I4I5 - I567	8	77I3I	77,2I
50	223	33I92	33,23	I568 - I779	7	78EI5	78,70
60	I90	3520I	35,24	I780 - 2044	6	80205	80,29
70	I74	36999	37,04	2045 - 2389	5	8I930	82,0I
80	I53	38634	38,67	2390 - 2980	4	84294	84,38
90	I35	40070	40,II	298I - 39I8	3	87I08	87,20
I00	I24	4I358	4I,40	39I9 - 5972	2	9I2I6	9I,3I
200	7I	50284	50,34	5973 -I4654	I	99898	I00,00
300	44	55785	55,84				

Таблица 3

Частотно-ранговое распределение лексики по данным ЧС словоформ авторской речи эстонской художественной прозы (i - ранг, P - частота, P^* - накопленная частота, P^* (%) - покрываемость текста)

i	P_i	P_i^*	P_i^*(%)	i	P_i	P_i^*	P_i^*(%)
I	322I	322I	3,22	200	48	35066	35,I0
2	I602	4823	4,83	300	34	38592	38,63
3	I439	6262	6,27	400	25	4I484	4I,53
4	I375	7637	7,64	500	2I	43755	43,80
5	I264	890I	8,9I	600	I7	45634	45,68
6	III6	I00I7	I0,03	700	I5	47264	47,3I
7	995	II0I2	II,02	800	I3	48703	48,75
8	7I3	II725	II,74	900	I2	49966	50,02
9	592	I23I7	I2,33	I000	II	5II09	5I,I6
I0	542	I2859	I2,87	I00I-I034	II	5I483	5I,54
20	329	I680I	I6,82	I035-II68	I0	52823	52,88
30	224	I9344	I9,36	II69-I299	9	54002	54,06
40	I89	2I384	2I,4I	I300-I500	8	556I0	55,67
50	I65	23I30	23,I5	I50I-I753	7	5738I	57,44
60	I4I	24635	24,66	I754-213I	6	59649	59,7I
70	I20	25920	25,95	2I32-2650	5	62244	62,3I
80	I08	27045	27,07	2651-3460	4	65484	65,55
90	89	28017	28,05	346I-5088	3	70368	70,44
I00	83	2887I	28,90	5089-8973	2	78I38	78,22
				8974-30733	I	99898	I00,00

Таблица 4

Значения параметров распределения Ципфа по данным ЧС эстонского языка *

Параметры	Лексемы	Словоформы
По формуле (I):		
C	6823	4I22
k	0,0683	0,0426
γ	0,92	0,86
По формуле (2):		
C	I4785	4595
k	0,I48	0,046
γ	I,04	0,87
β	2,3	0,5
Объем текста	99898	99898
Объем словаря	I4654	30733

* Значения параметров вычислены на основе средних рангов методом наименьших квадратов. Средние ранги вычисляются при образовании "платформ", т.е. когда встречается ряд слов одинаковой частоты. Например, по данным ЧС лексем (табл. I и 2) частоте $f = I$ соответствуют ранги $i = 5973 \div I4654$; средний ранг $i = I0300$. (О промежуточных вычислениях параметров на основе средних рангов см. Калинин В.М., I964, с. I25.)

Таблица 5

Значения параметра γ в разных частотных зонах по данным ЧС эстонского и русского языков (I – Засорина Л.Н., I966; II – ЧС русского языка, I977)

Частотная зона (i)	Эстонский язык		Русский язык (лексемы)	
	лексемы	словоформы	I	II
I ÷ 30	0,83	0,77	0,70	0,7I
30÷ 2500	0,98	0,89	0,94	0,95
> 2500	I,03	0,75	I,24	I,5I
Весь словарь	0,92	0,86	0,93	I.0
Объем словаря	I4654	30733	I0630	39268
Объем текста	99898	99898	I20474	I056382

Таблица 6

Значения параметра γ в разных частотных зонах по данным ЧС лексем I-го тома романа "Правда и справедливость" ("Tõde ja õigus") А.Х. Таммсаре (вычисления на основе данных: Villup A., I978)

Частотная зона (i)	Весь роман (I-й том)	В том числе	
		авторская речь	речь персонажей
I ÷ 30	0,7	0,68	0,59
30 ÷ I500	I,I	I,II	I,2I
> I500	I,4	I,43	I,47
Весь словарь	I,0	I,0	I,0I
Объем словаря	8228	7346	3I35
Объем текста	I60356	II4I24	46232

Таблица 7

Дробно-линейная модель. Наблюдаемые и ожидаемые частоты (F_i) по данным ЧС лексем авторской речи эстонской художественной прозы (на основе линейной связи между $1/i$ и $1/F_i$; ср. рис. 3). Вычисления по формуле:

$$F_i = (\alpha + \delta i)^{-1}$$

(или $F_i = C(i+B)^{-1}$, где $C = \delta^{-1}$ и $B = \alpha C$)

i	F_i (набл.)	$1/F_i$	f_i (ожид.)
I	4237	0,024 · 10^{-2}	4444
2	3493	0,029	3333
3	2598	0,038	2667
4	I98I	0,050	2222
5	I395	0,072	I904
I0	827	0,I2	IIII
50	223	0,45	256
I00	I24	0,8I	I30
500	27	3,70	27
I000	I3	7,69	I3
II00	II	9,09	I2
I200	I0	I0,00	II
I300	9	II,II	I0
I400	9	II,II	9
I500	8	I2,50	9
2000	6	I6,67	7
3000	4	25,00	4
5000	2	33,33	3
I0000	I	I00,00	I
Параметры:			
α		0,000I5	-
δ		0,000075	-
C		-	I3333
B		-	2

Таблица 8

Модель Арапова–Ефимовой–Прейдера ("дискретный аналог" закона Ципфа). Наблюдаемые и ожидаемые частоты (f_i) по данным ЧС авторской речи эстонской художественной прозы. Вычисления по формуле

$$f_i = \frac{B(L+1)^{\beta}}{1-\gamma}\left[(i+1)^{1-\gamma} - i^{1-\gamma}\right]$$

Ранг i	ЧС лексем		ЧС словоформ	
	f_i (набл.)	f_i (ожид.)	f_i (набл.)	f_i (ожид.)
I	4237	3I82	522I	3479
2	3493	I945	I602	2I97
3	2598	I4I8	I439	I636
4	I98I	II23	I375	I3I6
5	I395	932	I264	II06
I0	827	5I4	542	633
50	223	I2I	I65	I64
I00	I24	64	83	9I
500	27	I5	2I	23
I000	I3	8	II	I3
I500	8	6	8	9
3000	4	3	4	5
5000	2	2	3	3
I0000	I	I	I	2
20000	-	-	I	I
Параметры:				
L	I4654		30733	
γ	0,92		0,86	
β	0,6562		0,6608	

(N = 99898)

Л И Т Е Р А Т У Р А

Алексеев П.М. О нелинейных формулировках закона Ципфа. – В кн.: Вопросы кибернетики. Вып. 41. М.; Л., 1978, с. 53–65.

Алексеев П.М. Методика квантитативной типологии текста. Л., 1983. – 76 с.

Арапов М.В., Ефимова Е.Н., Шрейдер Ю.А. О смысле ранговых распределений. – Научно-техническая информация. Серия 2. М., 1975, № 1, с. 9–20; № 2, с. 9–20.

Арапов М.В., Ефимова Е.Н. Понятие лексической структуры текста. – Научно-техническая информация. Серия 2. М., 1975, № 6, с. 3–7.

Арапов М.В., Шрейдер Ю.А. Закон Ципфа и принцип диссимметрии системы. – В кн.: Семиотика и информатика. Вып. 10. М., 1978, с. 74–95.

Бернштейн Н.А. Очерки по физиологии движений и физиологии активности. – М.: Медицина, 1966.

Борода М.Г., Поликарпов А.А. Закон Ципфа-Мандельброта и единиц различных уровней организации текста. – Учен. зап. Тарт. ун-та, вып. 689. Труды по лингвостатистике. Тарту, 1984, с. 35–60.

Бычков В.Н. К проблеме обобщения и интерпретации ранговых распределений в статистической лингвистике. – Учен. зап. Тарт. ун-та, вып. 689. Труды по лингвостатистике. Тарту, 1984, с. 61–70.

Засорина Л.Н. Автоматизация и статистика в лексикографии. – Л.: Изд-во ЛГУ, 1966. – 128 с.

Калинин В.М. Некоторые статистические законы математической лингвистики. – Проблемы кибернетики. Вып. 11. М., 1964.

Козачков Л.С. Некоторые методологические вопросы теории информационно-поисковых систем. – Научно-техническая информация. Серия 2. М., 1969, № 12, с. 9–16.

Крылов Ю.К. Об одной парадигме лингвостатистических распределений. – Учен. зап. Тарт. ун-та, вып. 628. Труды по лингвостатистике. Тарту, 1982, с. 80–102.

Ланд К.Ч. Сравнительная статика в социологии. – В кн.: Математика в социологии. М.: Мир, 1977, с. 371–401.

Лебедев А.Н. Закономерности повторения слов в речи. – Психологический журнал, 1983, № 5, с. 11–22.

Мартыненко Г.Я. Некоторые закономерности концентрации и рассеяния элементов в лингвистических и других сложных системах. – В кн.: Структурная и прикладная лингвистика. Вып. 1. – Л.: Изд-во ЛГУ, 1978, с. 63–79.

Митропольский А.К. Техника статистических вычислений. Изд. 2-е. – М.: Наука, 1971. – 576 с.

Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. Вып. 1./ Перев. с англ. – М.: Финансы и статистика, 1982.– 320 с.

Орлов Ю.К. Обобщение закона Ципфа-Мандельброта. – Сообщения АН ГССР, т. 57, № 1, 1970, с. 37–40.

Орлов Ю.К. Обобщенный закон Ципфа-Мандельброта и частотные структуры информационных единиц различных уровней. – В кн.: Вычислительная лингвистика. М.: Наука, 1976, с. 179–202.

Орлов Ю.К. Модель частотной структуры лексики. - В кн.: Исследования в области вычислительной лингвистики и лингвостатистики. М.: Изд-во МГУ, 1978, с. 59-118.
Перебийніс В.С. Кількісні та якісні характеристики системи фонем сучасної української літературної мови. - Київ: Наукова думка, 1970. - 270 с.
Пиотровский Р.Г. Текст, машина, человек. - Л.: Наука, 1975. - 327 с.
Плат У. Математическая лингвистика. - В кн.: Новое в лингвистике. Вып. IV. М.: Прогресс, 1965, с. 201-245.
Туядава Ю. О квантитативных характеристиках богатства лексического состава художественных текстов. - Учен. зап. Тарт. ун-та, вып. 437. Linguistica. Тарту, 1977, с. 159-175.
Туядава Ю. К вопросу об аналитическом выражении связи между объемом словаря и объемом текста. - Учен. зап. Тарт. ун-та, вып. 549. Труды по лингвостатистике. Тарту, 1980, с. 113-144.
Хайтун С.Д. Наукометрия - состояние и перспективы. - М.: Наука, 1983. - 344 с.
Частотный словарь русского языка. / Под ред. Л.Н. Засориной. - М.: Русский язык, 1977. - 936 с.
Шрейдер Ю.А. О возможности теоретического вывода статистических закономерностей текста (к обоснованию закона Ципфа). - В кн.: Проблемы передачи информации. Т. 3, вып. I. М., 1967, с. 57-63.
Яблонский А.И. Стохастические модели научной деятельности. - В кн.: Системные исследования. Ежегодник 1975. М.: Наука, 1976, с. 5-42.
Hill B.M. A Theoretical Derivation of the Zipf (Pareto) Law. - In: Studies on Zipf's Law. / Quantitative Linguistics, vol. 16. Bochum: Brockmeyer, 1982, pp. 53-64.
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.
Kučera H., Francis W.N. Computational Analysis of Present-day American English. - Providence, R. I: Brown University Press, 1967. - 424 pp.
Mandelbrot B. Structure formelle des textes et communication: deux études. - Word, vol. 10, 1954, № 1, pp. 1-27.
Marx W., Schürer-Necker E. Überlegungen zur Interpretation des Zipfschen Gesetzes am Beispiel der frühen Kindersprache. - In: Glottometrika 1. / Quantitative Linguistics, vol. 1. Bochum: Brockmeyer, 1978, pp. 154-167.
Simon H.A. On a Class of Skew Distribution Functions. - Biometrika, vol. 42, 1955, pp. 425-440.
Villup A. A.H. Tammsaare romaani "Tõde ja õigus" I köite autori- ja tegelaskõne sagedussõnastik. - Учен. зап. Тарт. ун-та, вып. 446, Труды по лингвостатистике III, Тарту, 1978, с. 5-106.

Woronczak J. On an Attempt to Generalize Mandelbrot's Distribution. - In: To Honor Roman Jakobson. Vol. 3. The Hague: Mouton, 1967, pp. 2254-2268.
Zipf G.K. The Psycho-Biology of Language. Cambridge, Mass., 1935.
Zipf G.K. Human Behavior and the Principle of Least Effort. Cambridge, Mass.: Addison-Wesly Press, 1949.

THE STATISTICAL STRUCTURE OF TEXT AND ZIPF'S LAW

Juhan Tuldava

S u m m a r y

The statistical (or frequency) structure of text on the lexical level is considered to be a combination of two counterparts: the rank distribution and the spectral distribution of lexical units. In this article the rank distribution is examined and illustrated on practical material. The author gives a critical survey of some earlier and more recent versions of Zipf's law (G. Zipf, 1935; J. Woronczak, 1967; Yu. Orlov, 1970; V. Arapov, 1975; P. Alexeyev, 1978; Yu. Krylov, 1982; V. Bychkov, 1984) and discusses the problems of the theoretical substantiation of the law.

ЧАСТОТНЫЙ АНГЛО-РУССКИЙ СЛОВАРЬ ПО РОБОТОТЕХНИКЕ

Б.И. Шуневич

За последние десятилетия возник целый ряд новых отраслей техники, например, космонавтика, кибернетика, робототехника и другие. Это обусловило значительный рост терминологической лексики: возникновение новых терминов, а также расширение значений уже существующих терминов.

Одним из показателей интенсивного развития научно-технической терминологии в нашей стране и за рубежом является издание большого количества различных отраслевых словарей / Английский..., 1982; Иванова Э.Г., Филатов А.М., Хайлова И.М., 1979; Шишмарев А.И., Заморин А.П., 1978; De Vries Louis, 1978/ и публикаций, которые касаются принципов составления таких словарей / Алексеев П.М., 1975, 1983; Даниленко В.П., Скворцов Л.И., 1982; Дубровина Л.В., 1982; Иванов А.Н., 1971; Ступин Л.П., 1982; Brunner Th. T. and Berkowitz L., 1967 /. Двуязычные словари, в том числе словари новых терминов, необходимы для перевода и, следовательно, более широкого использования зарубежной научно-технической литературы специалистами, а также для улучшения контактов между научно-техническими работниками различных стран. Что касается робототехники, то, несмотря на ее бурное развитие и на огромную потребность в соответствующих справочных пособиях по этой отрасли техники, до сих пор не существует англо-русского словаря по робототехнике.

Для того чтобы помочь специалистам - робототехникам, переводчикам и всем лицам, интересующимся английской технической литературой по данной тематике, нами был составлен англо-русский словарь по робототехнике. Словарь составлялся с учетом частотности слов, что дает возможность использовать его не только для перевода, но и для выполнения различных лингвистических исследований, а также для отбора лексического минимума, необходимого преподавателям английского языка на специальных факультетах и отделениях вузов.

Прежде чем составить словарь, нужно было правильно организовать выборку с тем, чтобы она была репрезентативной и однородной. Репрезентативность требует, чтобы выборка включала все основные виды и жанры текстов и все композиционные части каждого текста, чтобы она по объему была достаточной

для получения статистически верных данных. Однородность предполагает однородность хронометрическую и тематическую.

Нами включены тексты с 1975 по 1979 годы, что позволяет считать выборку хронометрически однородной. Для соблюдения тематической однородности выделено 5 разделов для каждого из которых данные обрабатывались отдельно, а затем объединялись после проверки на однородность.

Для определения процентного количества информационных источников, относящихся к робототехнике, в различных видах научно-технической литературы, нами было выписано около 600 наименований этих источников из отечественных реферативных журналов, указателей переводов, экспресс-информаций и других. Оказалось, что из этих 600 источников информации по робототехнике 52,5 % относились к статьям из научно-технических журналов, 28,4 % - к материалам конференций, симпозиумов и 19,1 % - к патентным описаниям. Соответственно этому в выборку вошло различное количество текстов указанных трех массивов / см. таблицу I /. Определены и процентные соотношения между разделами робототехники внутри каждого из трех упомянутых массивов текстов и между пятью разделами робототехники для всего материала.

Проведенная работа дала возможность правильно организовать выборку для составления словаря, так как при подборе минимальных выборок / то есть текстов или их отрезков по 1000 словоупотреблений каждый / мы сохраняем уже установленные процентные соотношения между видами текстов научно-технической литературы и соответственно разделам робототехники.

Словарь был составлен вручную на материале 211 минимальных выборок. Распределение их по массивам текстов и разделам представлено в таблице I.

В выборке представлены тексты по робототехнике из основных англоязычных стран, например: США представлены 40 патентными описаниями и 22 статьями из различных научно-технических журналов, Великобритания - 10 патентами и 66 статьями из различных научно-технических журналов, Канада - 4 статьями из одного журнала, Австралия - 3 статьями из одного журнала. Кроме этого использованы 57 докладов и сообщений, прочитанных на пяти международных конференциях и симпозиумах по данной тематике.

Таблица I

Распределение выборок по разделам подъязыка робототехники

№ п/п	Вид литературы / Название раздела	Патенты		Труды конферен.		Журн. статьи		Всего	
		колич.	%	колич.	%	колич.	%	колич.	%
I	Общие проблемы	-	-	5	9	22	2I	27	I3
2	Конструкции, целевые механизмы, узлы оснастка роботов и манипуляторов	34	68	4	7	I6	I5	54	26
3	Исследования, разработки моделей роботов и манипуляторов	-	-	I0	I7	9	9	I9	9
4	Системы управления роботов и манипуляторов	8	I6	25	44	I7	I6	50	24
5	Применение роботов и манипуляторов	8	I6	I3	23	40	39	6I	29
Всего		50	I00	57	I00	I04	I00	2II	I00

Примечание: В таблице указаны краткие названия разделов робототехники.

Как видно из таблицы, выборки подбирались из трех видов научно-технической литературы: I/ патентных описаний - 50; 2/ трудов конференций - 57; 3/ журнальных статей - I04.

Большинство текстов, взятых для исследования, были значительно длиннее, чем I000 словоупотреблений. В этих случаях для обеспечения репрезентативности компонентного строения полных текстов выборки брались из различных частей научно-технических произведений. Например, при общем количестве 57 выборок из трудов конференций и симпозиумов I9 из них извлечены из начальных частей текстов, I9 из средних и I9 из конечных.

Распределение 600 названий текстов по разделам осуществлялось при помощи специалистов по робототехнике, а количественный и качественный подбор текстов / 2II названий /, из которых взяты выборки - с учетом соответствующих рекомендаций группы " Статистика речи ", работающей под руководством проф. Пиотровского Р.Г. / Алексеев П.М., I975, I983 /.

Как оказалось, процентное соотношение этих разделов и даже их количество зависит от массива текстов: так, в патентных описаниях самый большой вес имеет 2 раздел / 68 % текстов патентных описаний /; в трудах конференций - 4 раздел - 44 %; в журнальных статьях - 5 раздел - 39 %.

Определение единиц подсчета осуществлялось следующим образом:

I/ термин определялся по словарю лингвистических терминов / Ахманова О.С., I969, с. 474 /;

2/ слова, написанные через дефис, считались одним словом, например: point - to - point - позиционный, range-finder - дальномер;

3/ одна аббревиатура или сокращение считались одним словом, например: CNC / computer numerical control / - ЧПУ /числовое программное управление /, pot / potentiometer - потенциометр;

4/ математические формулы, числительные, написанные цифрами, при подсчете длины выборки в рассчет не принимались.

Из анализируемых текстов выписывались все однословные термины и многокомпонентные термины по робототехнике с препозитивными и постпозитивными определениями / типа N prp N /, а также наиболее часто употребляемые общетехнические и общенаучные термины, например: arm, device, industrial robot, degree of freedom.

Все термины в словаре стоят в исходной форме, то есть имена существительные в единственном числе, кроме существительных, употребляемых только во множественном числе / Pluralia Tantum /, глаголы в инфинитиве / за исключением Participle I и Participle II, так как они играют особую роль при образовании терминов и считаются самостоятельными классами/.

Сочетания, в которых при одном ядре есть несколько определений, соединенных союзами and, or, рассматривались как объединение нескольких терминов, если сочетание ядра с каждым определение встречалось в текстах как самостоятельная единица,

например, терминологические сочетания:

- arm or wrist joint = arm joint + wrist joint,
- program and teaching facilities =
  program facilities + teaching facilities,
- yaw, pitch and roll motions = yaw motion + pitch motion + roll motion.

Некоторые словосочетания с союзом and представляли собой не сумму терминов, а части одного термина, не употребляющиеся самостоятельно, например, многокомпонентные термины:

- rack and pinion arrangement ≠
  rack arrangement + pinion arrangement,
- pick and place robot ≠ pick robot + place robot

Составленный нами частотный англо - русский словарь содержит около 7 тысяч лексических единиц и состоит из двух частей и четырех приложений: часть I - алфавитно - частотный словарь однословных и многокомпонентных терминов по робототехнике и наиболее часто встречающихся общенаучных и общетехнических терминов; часть II - частотные списки однословных и многокомпонентных терминов, начиная с наибольшей частоты и кончая единичными употреблениями. В приложения включены: список сокращений, таблица моделей многокомпонентных терминов, которые выявлены в патентных описаниях, трудах конференций и журнальных статьях, таблицы распределения частот и рангов однословных и многокомпонентных терминов.

Заголовочные единицы английской части словаря - это термины в общепринятом понимании и слова, входящие в состав многокомпонентных терминов в качестве опорных элементов.

Каждая словарная статья строится следующим образом: английское слово; указания на часть речи, к которой оно относится; число, обозначающее частоту слова в текстах; русский перевод.

Словарь построен по алфавитно - гнездовой системе. Поэтому термины, состоящие из определений / адъюнктов / и определяемых слов / ядер /, следует искать по определяемым словам.

Например, многокомпонентный термин continuous path robot следует искать в гнезде термина robot. Ядро термина в гнезде заменяется тильдой / /. При терминах также указываются частоты их употребления в общей выборке. Например:

robot n	I483	робот
adaptive	3	адаптивный робот
assembly	I5	робот для сборки, сборочный робот
computer-controlled	6	робот, управляемый от ЭВМ
continuous-path	2	робот с контурной системой управления
electromechanical	3	электромеханический робот
industrial	292	промышленный робот
intelligent	9	интеллектуальный робот
painting	5	окрасочный робот, робот для окраски
pneumatic	4	пневматический робот, робот с пневмоприводом
programmable	8	программируемый робот
welding	I4	сварочный робот, робот для сварки
robotics n	5	робототехника
robotize v	2	роботизировать

В процессе работы над словарем были выявлены все модели многокомпонентных терминов, которые встретились в каждом из вышеупомянутых трех массивов текстов и во всем массиве текстов, а также выделены те, которые наиболее частые в текстах по робототехнике. Ими являются: NN, AN, VN, NNN, ANN, NNNN, ANNN, N of N . Например: **control system, hydraulic robot, clamping range, master - slave manipulator, numerical control library, image - storage inspection system, binary input-output line, degree of freedom.**

Ниже приводятся фрагменты частотных списков однословных и многокомпонентных терминов для общей выборки / начальные части при f 63 и f 39 соответственно /.

В первое приложение включены термины-аббревиатуры с указанием их абсолютных частот, дана расшифровка и перевод их на русский язык.

Во втором приложении приводятся модели многокомпонентных терминов в трех массивах текстов с указанием их абсолютных частот в каждом массиве текстов отдельно, а также сводная таблица для трех массивов текстов.

В третьем и четвертом приложениях-таблицы распределения

Частотный список однословных терминов
по робототехнике/при $f > 63$ /

1	robot n	1483	39	task n	149
2	part n	540	40	information n	144
3	position n	427	41	figure n	143
4	arm n	418	42	control v	140
5	use v	367	43	signal n	138
6	provide v	363	44	body n	133
7	operation n	347	45–46	assembly n, com-	
8	system n	343		ponent n	132
9	object n	322	47	rotation n	130
10	movement n	320	48	form n	127
11	machine n	318	49	tool n	126
12	axis n	306	50	value n	124
13	application n	302	51–52	area n, cylin-	
14	move v	287		der n	123
15	point n	249	53	sequence n	122
16	manipulator n	248	54–55	design n, member n	121
17	program n	246	56	operate v	120
18	hand n	235	57	housing n	118
19	means n	232	58	process n	115
20	workpiece n	230	59–60	invention n,	
21	type n	227		input n	114
22	direction n	223	61–62	surface n, welding n	113
23	end n	222	63–64	level n, shaft n	111
24	use n	218	65	step n	110
25	line n	207	66–67	develop v, job n	109
26	finger n	200	68–71	describe v,	
27	control n	192		method n, model n,	
28	time n	189		wrist n	107
29–30	computer n, motion n	186	72–73	cost n, plate n	106
31	joint n	174	74–75	memory n, set n	105
32	problem n	173	76–77	orientation n,	
33	function n	171		speed n	104
34	require v	163	78–80	development n,	
35	sensor n	161		device n,	
36	operator n	160		range n	103
37–38	case n, gripper n	156	81	force n	102

82	feature n	98
83	motor n	97
84-85	data n, programming n	95
86-87	requirement n, station n	94
88	portion n	93
89	capability n	92
90-91	angle n, design v	91
92	using n,a	89
93-94	claim n, path n	87
95-96	mount v, space n	86
97-98	co(-)ordinate n, handling n	84
99-100	produce v, result n	83
101-102	cycle n, determine v	81
103	base n	80
104-105	output n, side n	79
106-108	company n, distance n,location n	78
109-110	equipment n,press n	77
111-112	connect v,plane n	76
113-114	action n, store v	75
115-116	group n, hole n	74
117	human a	73
118	user n	72
119- 122	command n, conveyor n, form v, element n	71
123-125	equation n,lenght n, mounted a	70
126-130	block n, example n, handle n, parallel a, section n	69
131	controller n	68
132-134	ability n, configuration n, installation n,	67
135-139	basis n, environment n, mode n, program n, term n	66
140-143	moving n, size n, standart n, tube n	65
144-147	condition n, controlling n,a, module n, table n	64

Частотный список многокомпонентных терминов по робототехнике / при f > 39 /

1	industrial robot	292
2	degree of freedom	145
3	control system	113
4	machine tool	67
5	co(-)ordinate system	65
6	robot arm	53
7	manipulator arm	40

частот и рангов однословных и многокомпонентных терминов для трех массивов текстов.

Словарь предназначается для научных и инженерно - технических работников в области робототехники, переводчиков, преподавателей, студентов и для перевода специальной литературы, для исследований в области лингвистической статистики и статистической лексикографии, а также для оптимизации преподавания иностранных языков.

Л И Т Е Р А Т У Р А

Алексеев П.М. Статистическая лексикография / Типология составления и применения частотных словарей /: Учебное пособие / П.М. Алексеев; Ленингр. гос. пед. ин - т им. А.И. Герцена. - Л.: ЛГПИ, 1975.

Алексеев П.М. Методика квантитативной типологии текста: Учебное пособие. / П.М. Алексеев; Ленингр. гос. пед. ин - т им. А.И. Герцена. - Л.: ЛГПИ, 1983.

Английский частотный словарь по кибернетике./ Сост. Р.В. Адамов, Л.Н. Александрова, В.М. Андрющенко и др.; Предисл. М.М. Глушко, Л.Н. Александровой /; Под ред. М.М. Глушко. - М.; Изд - во МГУ, 1982.

Ахманова О.С. Словарь лингвистических терминов. - М.: Сов. энциклопедия, 1969.

Даниленко В.П., Скворцов Л.И. Лингвистические проблемы упорядочения научно - технической терминологии. - ВЯ, 1982, №1.

Дубровина Л.В. Англо - русский медицинский терминологический ключ./ Методика составления терминологического словаря /. - М.: МГУ, 1982

Иванов А.Н. Пополнение словарного состава современного английского языка./ Опыт лексикологического и социолексикологического описания /. АКД. М., 1971.

Иванова З.Г., Филатов А.М., Хайлова И.М. Японско - русский и русско - японский словарь по системотехнике и робототехнике. - М.: Русский язык, 1979.

Ступин Л.П. Теория и практика английской лексикографии.-Л.: ЛГУ, 1982.

Шишмарев А.И., Заморин А.П. Англо - русско - немецко - французский толковый словарь по вычислительной технике и обработке данных. Под ред. акад. Дороницына А.А. - М.: Русский язык, 1978.

Brunner Th. T. and Berkowitz L. The elements of scientific and specialized terminology. - Minneapolis, 1967.
De Vries Louis German - English science dictionary. - New York, Mc. Graw - Hill, 1978.

ENGLISH-RUSSIAN FREQUENCY LIST ON ROBOTICS

Bogdan Shoonevich

S u m m a r y

The article deals with the principles of material selection for compiling an English-Russian frequency list on robotics; its structure is described; patterns of most frequently used multi-componential terms in patents specifications, in journal articles and conference proceedings in the given sphere, as well as in the general selection including all the above-mentioned texts are given.

ИЗ ИСТОРИИ ЛИНГВОСТАТИСТИКИ

ЭВФОНИКА "НЕЗНАКОМКИ"[+]

А.М. Финкель (1899–1968)

I.

Звуковая организация стихотворного произведения отнюдь не является несущественной или маловажной. Наоборот, она вплетается в общую систему художественных средств с неменьшей силой и выразительностью, чем все остальные компоненты. Однако характер звуковых средств, способ их организации, их выразительность и эстетическое воздействие вовсе не такого рода, как у лексики или синтаксиса. И первое, от чего здесь следует отрешиться, это от увязки – явной или тайной – звучания с идеями. Конечно, полное уравнение организованного словесного звучания с звучанием музыкальным тоже себя не оправдывает, ибо эти две фонетических системы не тождественны. Законы музыкальной гармонии и музыкальной инструментовки на искусство слова не распространяются, и термины эти по отношению к литературе имеют характер метафорический, т.е. условный, неточный и с прямым значением не тождественный. На звучание слова нельзя переносить даже таких широких понятий, как мажор или минор, не говоря уже о множестве других, более специфических. И в то же время, поскольку в слове звуки неразрывно связаны со значением, ассоциации, устанавливаемые для одного какого-то смыслозвукового единства, переносятся в какой-то мере с наиболее знакомого слова и на звучание соседних слов. В результате этого единство (или близость) звучания влечет за собой по ассоциации и некоторое единство смысловое и эмоциональное. Сила этих ассоциаций далеко не одинакова и часто равна нулю. Различную силу ассоциативных связей между звуками и значением можно показать на следующих примерах.

[+] Подготовка текста и публикация С. И. Гиндина. Необходимые поправки и дополнения к отдельным положениям автора даны в помещенных в конце статьи "Примечаниях публикатора". Характеристика источников публикации и общая оценка работы А.М. Финкеля содержатся в публикуемой ниже статье С.И. Гиндина.

В известных стихах К. Бальмонта: Вечер. Взморье. Вздохи ветра, величавый возглас волн - звук В[1], несмотря на настойчивое его повторение, никаких смысловых или эмоциональных обертонов не вызывает, как не вызывает их и повторение звука Ч в строке Чуждый чарам черный челн[2] - их повторение ничем не оправдано. В то же время вполне оправдано художественное повторение звука Ш в такой, к примеру, строке: Шелест шелка, шум и шорох в мягких пурпуровых шторах (Э.По. Ворон)[3], ибо связь звука Ш со значением слов шелест, шорох и шум несравненно сильней и прочней, чем у звука В со значением слов вечер или возглас (из-за наличия или отсутствия звукоподражания).

Существует и звукопись, и словесная инструментовка, однако в изучении их следует соблюдать максимальную трезвость и осторожность, избегая психологизма и субъективизма и не приписывая самим явлениям своих представлений о них.

Говоря о фонетической организованности "Незнакомки", мы имеем в дальнейшем в виду явления двоякого рода: с одной стороны, то, что можно назвать "микрозвфонией" и что не выходит за пределы одной строфы - это так называемые аллитерации, ассонансы и т.п., а с другой, - те явления, которые охватывают большие отрывки, включая сюда как общую характеристику звучания, так и те звуковые повторы, которые при помощи повторяющихся слогов, слов и словосочетаний связывают целые строфы, выполняя тем самым композиционное задание.

2.

В каких случаях, с какого момента мы получаем право говорить о какой-то специфической, организованной, упорядоченной эвфонии поэзии в отличие от хаотического, неупорядоченного звучания разговорной речи, письменного нехудожественного языка и даже, может быть, художественной прозы? На это пытался дать ответ А.М. Пешковский в своей интересной методологически и полезной фактическими данными статье (Пешковский, 1925).

Основные предпосылки А.М. Пешковского, всецело нами разделяемые, таковы.

Прежде чем говорить о так называемой звуковой инструментовке или звукописи, то есть о каком-то индивидуальном подборе или расположении звуков, прежде чем приводить это в связь то с содержанием произведения, то с общими законами благозвучия, исследователю следует иметь хоть какую-нибудь объективную мерку. Без нее он не может знать, какова обыч-

ная, нецеленаправленная частота употребления данного звука в речи, а потому на каждом шагу рискует впасть в ошибки, не заметить одних фактов, переоценить другие и т.д. Работа Пешковского и является попыткой установить на материале 10 000 звуков обычной разговорной речи некотрые объективные показатели, характеризующие звучание русского языка. Мы не будем излагать здесь, какую методику записи и вычислений избрал А.М. Пешковский, не будет подвергать критике точность его приемов, - мы принимаем его данные в качестве той объективной мерки, по отношению к которой производится сравнение фонетической структуры "Незнакомки". Но для того, чтобы это сравнение было показательным, нам пришлось применить ту же методику, что и А.М. Пешковский, даже если не во всем мы с ней были согласны. Таким образом, звучание поэтического произведения сопоставляется с "нормой" звучания русской литературно-разговорной речи, что и позволяет сделать некоторые небезинтересные выводы. Как писал А.В. Федоров, "Организация звуковых качеств является в большом числе случаев моментом лишь сопутствующим, а иногда вовсе не играет никакой роли. ... говорить об "инструментовке", фонической организации, значении звука, как такового, - можно в том случае, если есть установка на звуковой фактор, если организованность звуковых качеств в пределах известных единиц и единств расчитана и воспринимается, как факт структурный, т.е. в том случае, когда наличествуют повторы, обусловленные не только чисто языковыми условиями (например, частотою данного звука в языке)" (Федоров, 1928, с. 57).

3.

Сравнение фонетической структуры "Незнакомки" с данными Пешковского показывает, что в отношении многих сторон существенного различия между обычной разговорно-литературной речью и данным стихотворением нет.

Если сравнить соотношение г л а с н ы х и с о г л а с н ы х, то никакой заметной разницы мы не обнаружим: в разговорно-литературной речи гласные составляют 42,3 %, а у Блока 41,4 %; согласные, соответственно, 57,7 и 58,6 %. Различия эти настолько малы, что их можно не принимать во внимание. Поскольку же звучность речи зависит прежде всего от количества гласных и прямо ей пропорциональна, то первый вывод, который можно и надлежит сделать, состоит в том, что з в у ч н о с т ь " Н е з н а к о м к и " в о б щ е м н е о т л и ч а е т с я о т н о р м а л ь -

ной звучности русского языка.

Не отклоняется от нормы и соотношение между гласными ударными и безударными. По подсчетам А.М. Пешковского, их соотношение таково: 37 % ударных и 63 % безударных, то есть I:I,7. В "Незнакомке" эти отношения составляют 35 % и 65 %, то есть I:I,8. А так как ударяемый гласный - это гласный сильный, то отсюда следует, что и по силе звучания стихотворение Блока не отступает от нормальной литературно-разговорной речи.

Однако звучность речи зависит не только от соотношения гласных и согласных, ударяемых и неударяемых звуков, а еще и от качества гласных и согласных. Как известно, наиболее звучными из гласных являются гласные нижнего подъема (А), за ними следуют гласные среднего подъема (О,Е), и наименее звучны гласные высокие (И, Ы, У). Из согласных наиболее звучными являются сонорные (Р, Л, М, Н), за ними следуют шумные звонкие и, наконец, шумные глухие, которые даже называются безголосыми (atonae).

Сравнение "Незнакомки" с данными А.М. Пешковского показало, что наиболее звучные гласные распределены примерно одинаково: в разговорно-литературной речи они составляют 25,5 % всех гласных, а у Блока 24,0 %, что составляет весьма небольшое расхождение. Однако наименее звучные гласные (высокие, закрытые) у Блока занимают чрезвычайно большое место: 47 % при обычных для разговорно-литературной речи 38 %. Соответственно понижается удельный вес гласных средней звучности (по Пешковскому - 36,5 %, у Блока - 29 %). Таким образом, мы вправе сделать вывод, что в отношении гласных "Незнакомка" звучит более приглушенно, чем обычная речь. Понижение звучности гласных произошло в результате известного нам Ч частого повторения слова и: при нормальной частоте звука И в 5,3 % частота его в "Незнакомке" составляет 8,4 %.

Иначе звучат согласные. В то время, как в разговорно-литературной речи сонорные составляют лишь 28,4 %, а шумные - 71,6 %, в "Незнакомке" вес сонорных повышается до 34,2 % с соответственным понижением шумных. Но и шумные у Блока звучнее, чем в обычной речи: в ней звонкие составляют 38,3 % от общего числа шумных согласных, а у Блока 43,8 % (с соответствующим понижением глухих). Таким образом, с о-

гласные у Блока намного звучнее, чем обычно.

Если же взять весь звуковой состав "Незнакомки" в целом, то картина будет такова (в %%):

	По Пешковскому	У Блока
Гласные	42,3	41,4
Согласные сонорные	16,4	20,1
Согласные шумные:		
звонкие	15,8	16,9
глухие	25,5	21,6
Итого	100	100

Это значит, что по своему фонетическому составу стихотворение Блока звучнее, чем обычная разговорно-литературная речь, причем повышение звучности идет не за счет гласных, а за счет согласных. Можно даже указать, за счет каких именно согласных достигается эта звучность. Это, главным образом, звуки М, Н и Р, а их шумных - звонкий - Й. Сравнительные частоты их таковы (в %%):

Звуки:	М	Н	Р	Й
По Пешковскому	2,7	6,2	3,7	4,1
У Блока	3,7	8,2	4,8	5,6

Небезынтересно при этом отметить, что глухой шумный Т, частота которого в обычной разговорно-литературной речи составляет 7,6 %, у Блока встречается чуть ли не вдвое реже - 4,0 %.

И все же эту повышенную звучность языка "Незнакомки" нельзя считать чертой индивидуальной, присущей лишь данному произведению.

В своей статье А.М. Пешковский приводит свой подсчет глухих в первой тысяче звуков "Евгения Онегина" (Пешковский, 1925, с. 181). В то время, как в обычной разговорно-литературной речи глухие составляют 25,5 %, в "Евгении Онегине" они составляют лишь 20 %. Как формулирует А.М. Пешковский, разговорная речь "шумнее", а "Евгений Онегин" "звучнее" в этом пункте на 5,5 %. Но как только что было показано, и "Незнакомка" в этом пункте звучнее, если не на 5,5 %, то на 4 %, то есть стоит по своей звучности ближе к "Евгению Онегину", чем к разговорной речи. Возможно, что эта звучность вообще присуща русскому поэтическому, стихотворному языку, но это требует тщательной и точной статистической проверки.

Пока же мы считаем себя вправе сделать лишь один вывод: чем обширней исследуемый материал, тем сильнее проявляется на нем нивелирующее действие больших чисел. Поэтому некое художественное задание фонетика стиха может выполнять лишь на малых отрезках речи, а не в масштабе всего произведения в целом (разве что это произведение не превышает вообще 8–12 строк и не является канонической формой вроде сонета, триолета и т.п.).

То же следует сказать и в отношении так называемой звукописи, инструментовки, звуковых повторов. О них можно говорить лишь тогда, когда перед нами явное и резкое отклонение от нормы, но и это может проявиться лишь на небольшом отрезке. Приведем показательный пример. Звук В в обычной разговорно-литературной речи имеет (по Пешковскому) частоту в 3,6 %, а у Блока – 3,5 %; звук Э, соответственно, 3,4 % и 2,7 %. Отсюда, казалось бы, мы должны сделать вывод, что ни о какой звукописи здесь говорить не приходится, так как у Блока частота этих звуков ниже нормальной. Но в 9-й строфе "Незнакомки" в стихе И веют древними поверьями ... на 10 гласных приходится три Э (причем все они ударяемые, сильные), что составляет 30 %, а на 13 согласных (включая сюда и графически не обозначенный Й) – три В, что составляет 23 %. Это настолько превышает норму, что не обратить на это внимания или считать этот факт случайностью никак нельзя.+

Из всего этого следует, что от суммарных характеристик всего стихотворения в целом нужно перейти к рассмотрению отдельных его частей, ибо только в них может проявиться организация звукового материала. Рассмотрим прежде всего зву-

---

+ Конечно, для того, чтобы обнаружить эти повторяющиеся Э и В, никакой статистики не нужно. Как остроумно сказал А.М. Пешковский, "человек, подавший в зачумленный край, ... справедливо заметил бы без всякой статистики, по одним частым встречам с похоронами, что дело неладно" (Пешковский, 1925, с. 168). Статистика нужна для другого – для той объективности, без которой не может быть никакой науки. И об этом у Пешковского сказано очень хорошо. Образцом же субъективизма может служить известная (и исторически интересная) статья О. Брика (Брик, 1919), где автор усматривает повторы в любых сочетаниях, не учитывая нормальной частоты и повторяемости звуков[5].

звучание отдельных строф.

5.

Строфа первая отличается своей высокой звучностью. Звучность эта создается совместным действием двух наиболее звучных звуков: гласного А и согласного Р. Если звук А в этой строфе занимает то же место, что и в обычной речи и в "Незнакомке" вообще, то частота звука Р значительно выше: при норме в 3,7 %, а в "Незнакоме" в целом 4,8 %, в этой строфе частота звука Р достигает 8 %. Важно при этом и то, что Р и А действуют не разрозненно, а как единое целое, образуя повторяющееся несколько раз сочетание РА: по вечеРАм, над рестоРАнами, гоРЯчий, пРАвит, – сочетание вдобавок во всех случаях ударяемое, т.е. сильное и заметное.

Звучность этой строфы характеризуется еще и тем, что в ней количество гласных не намного уступает количеству согласных: 36 гласных на 52 согласных, т.е. отношение равно I:I,4 или 4I % и 59 %. И хотя в этой строфе звук И употреблен даже несколько чаще, чем А (II,4 %), вдвое превышая норму (5,3 %), однако из десяти И девять является безударными, причем семь из них даже заударными, т.е. глухими, редуцированными. Таким образом, доминирующим является открытый, широкий звук А. Если прибавить к нему звучание трех О, а к семи Р прибавить шесть Н, четыре М, два Л и четыре Й, то вся строфа предстанет перед нами как чрезвычайно звучная и широкая, намного превышающая звучность обычной разговорно-литературной речи (по взятым показателям 4I % против 35 %).

По-иному звучит вторая строфа. В ней также I2 ударных слогов, но ударения распределены почти равномерно между различными гласными (три А, три У, два Э, два И), так что ни о какой концентрации говорить не приходится. Вдобавок под ударением стоят звуки узкие, закрытые – И и У, т.е. менее звучные, чем А или О. Чрезвычайно большое место занимает редуцированный звук Ъ:II из 36, т.е. 30 % (при обычном его весе среди гласных I9,5 %, а в "Незнакомке" вообще даже I6 %). Общее отношение гласных к согласным равно 36:58, т.е. I,0 к I,6, или 38 % к 62 %. Всем этим создается звучание закрытое, приглушенное. Это усиливается и звучанием согласных, среди которых сонорные составляют лишь 29 % против 32% в первой строфе.

Третья строфа вновь возвращает нас к звучанию первой: такое же соотношение гласных и согласных (36:47 = I,0:I,3),

или 43 % и 57 %, такое же изобилие звука А (12 из 36, причем 3 из них ударяемые), такое же обилие сонорных и Й (19 из 47, т.е. 40 %), т.е. такая же открытость, сила и звучность.

Четвертая же строфа больше приближается ко второй: звуки высокого подъема преобладают над всеми другими, звук А находится под ударением всего один раз, а общая частота его не превышает нормальной. Звучность согласных также ниже, чем в предыдущей строфе. Отношение гласных к согласным равно I,0:I,6 (36:57).

6.

Иначе окрашена звучность пятой строфы. Общее соотношение гласных к согласным показывает еще больший перевес последних - 36 и 59, т.е. I:I,6. Однако ударяемых, т.е. сильных слогов в ней больше - I3 из 36, и под ударением находятся по преимуществу А (4) и О (4), т.е. гласные заднего ряда, широкие, низкого тона. Это звучание поддерживается еще звучанием семи А в первом предударном слоге, т.е. в позиции, в которой они мало отличаются от ударяемых. Таким образом, из 36 гласных открытые составляют I5, средние I5, а закрытые 6.

Что касается согласных, то здесь бросается в глаза обилие носовых (М и Н) - I3 из 59; многие из них находятся в ударном слоге, своеобразно окрашивая предшествующий гласный: еди́нственный, мое́м, стака́не, отраже́н, таи́нственный, смире́н, оглуше́н. Это подкрепляется вдобавок внутренней рифмой в последнем стихе (смирён - оглушён).

Трижды повторяется в 3-й строке краткое, обрывающееся окончание слов - влагой терпкой и таинственной, где конечное ЫЙ как бы отсекает одно слово от другого и обрубает каждое в отдельности.

Шестая строфа по составу гласных напоминает первую. В ней то же количество ударяемых А (5) и О (3), почти то же количество И (II против I0). Это создает одинаковую тембральную окраску гласных. Однако согласные звучат иначе: при почти одинаковом количестве сонорных и Й количество глухих в этой строфе заметно больше, чем в первой, и составляет 40 % всех согласных звуков (против 33 % в первой строфе) с соответствующим понижением роли звонких (I5 % против 23 %). Отношение гласных и согласных равно здесь I:I,3.

7.

Начиная с седьмой строфы стихотворение, как нам извест-

но[6], переходит в другой план, и отсюда начинается иная лексика, иная семантика, иной синтаксис. Эта новая тональность, изменение ключа, сопровождается и иным звучанием – более широким, более плавным, более звонким.

Если общее отношение гласных и согласных в седьмой строфе еще не очень отличается от того, что было раньше (оно такое же, как во второй или пятой), то характер самих гласных уже далеко не тот же самый. Из обычных 36 гласных ударяемых в этой строфе 15, что составляет 42 % и превышает средний процент их по всему стихотворению на 7 %. Если же учесть, что звук А составляет в данной строфе 25 % всех гласных, то звучание этой строфы определяется как чрезвычайно звучное и притом низкое, построенное на сильных и долгих открытых звуках. Столь же высок и вес сонорных – свыше 1/3 (20 из 59), что еще больше повышает звучность этой строфы. На фоне глухих шумных согласных (25 из 59) это звучание гласных и сонорных выделяется весьма рельефно.

Восьмая строфа продолжает это же звучание. Из 12 ударных слогов 9 включает звук А, образуя сквозное звучание – пройдя́ – пья́ными – всегда́ – одна́ – дыша́, духа́ми и т.д., а между этими широкими и низкими А вкрапляются одиночные Э, И, У. Так образуется ровное волнообразное движение сильных и звучных гласных, сопровождаемое весьма краткими, отрывистыми, высокими заударными И: пья́ными, духа́ми, тума́нами.

Количество согласных в этой строфе намного меньше, чем в предыдущей, – всего 45, т.е. 55,5 % общего количества при среднем для всего стихотворения 58,6 %. Вдобавок в этой легкой строфе 16 согласных (т.е. свыше 1/3) относятся к сонорным, из которых половину составляет звук Н, опирающийся на гласный А.

Еще более легкой является строфа девятая. В ней всего 80 звуков, и количество гласных немногим уступает количеству согласных (36 и 44), т.е. 45 % и 55 %. Одна треть гласных находится под ударением, и ударения эти распределены равномерно между нелабиализованными А и Э (по 4) и лабиализованными У и О (по 2). Поскольку лабиализация сопровождается понижением тона гласных, а звук А также низкий, то все звучание этой строфы проходит в низком тоне. На этом низком фоне и выделяется замечательный повтор И веют древними поверьями, дающий своим высоким звучанием почти физическое ощущение дуновения.

Сонорные составляют в этой строфе 36 % всех согласных,

причем преобладают Р и Л (10 из 16), что еще более увеличивает звучность этой строфы.

Дальше уже до самого конца стихотворения все время слышится эта обильная звучность, проявляющаяся заметней всего в почти равномерном распределении гласных и согласных (1:1,3).

8.

В д е с я т о й с т р о ф е доминирующим ударным гласным является звук О (4), сопровождаемый звуком У (1 ударяемый и 7 неударяемых). Будучи звуками заднего ряда и притом лабиализованными, они создают низкий тон звучания, поддержанный в свою очередь десятью низкими А (3 ударяемых + 7 неударяемых) и окрашивающий весьма своеобразно всю строфу.

Среди согласных половину, и даже несколько больше, занимают Й и сонорные, среди которых выделяется четырежды повторенное долгое Н. Количество глухих в этой строфе минимальное – всего 10. Это единственная строфа, в которой звонких больше, чем глухих, что и придает ей характерное звучание.

Строфа о д и н н а д ц а т а я является также одной из наиболее звучных и воздушных. И в ней гласные и согласные распределены в отношении 1:1,3, однако ударяемых гласных в ней больше – 15, что придает звучанию отчетливость и внятность. Но звучание здесь иное: максимум ударений приходится здесь на гласные полузакрытые – О (4) и Э (5), а гласные открытые занимают весьма скромное место. Особую окраску придает звучание рифм – на лабиализованные гласные О и У, что, как известно, приводит к понижению тона.

Среди согласных и здесь выделяются сонорные и Й, составляющие свыше половины их общего количества, – 25 из 48.

Двенадцатая строфа сохраняет то же отношение гласных и согласных. Ударяемые А (3), О (3) и У (3) с рифмами, построенными на лабиализованных звуках, сохраняют известное нам низкое звучание. Заметную опору ударяемые гласные находят в звуке Н: такие сочетания, как склоненных, бездонные, в которых соединены долгий гласный и долгий согласный, или же такое, как перья страуса, где ударный гласный подкреплен плавным Р, создают большую звучность, т.е. высокую воспринимаемость. Распределение согласных также весьма равномерно: 17 сонорных, 14 звонких и 17 глухих, не давая заметного преимущества ни одной из этих групп, образуют ровный, отчетливый, звучный тон.

Заключительная тринадцатая строфа отличается прежде всего большим весом ударяемых гласных: на 36 слогов их прихо-

дится I4, т.е. 40 %, что превышает среднее их количество на 5 %. Ударения распределены почти равномерно между звуками передними (6) и задними (8).

Что касается согласных, то здесь преобладают глухие; сонорные же занимают место, наиболее низкое во всем стихотворении (I4 при среднем в I7,5). Это значит, что хотя строфа в целом не уступает по звучности другим (вспомним обилие гласных), однако характер ее иного свойства, чем ранее, построенный на иных тембрах, на гласных полуоткрытых, на согласных глухих, т.е. в среднем регистре.

Строфа эта как бы примиряет взлеты и падения всех предшествующих и разрешает все это в последней строке, когда после широкого А (я зна́ю) следует многократно повторяющееся после мягких согласных переднее И и заключительное Э (и́стина в вине́), что создает в совокупности высокую тональность стиха.

9.

Какие из этого можно сделать в ы в о д ы ?

Как уже указывалось выше, в стихотворении Блока явственно обозначаются две части: первая (строфы I-6), написанные в "реально=достоевском" тоне, и вторая, в которой дается преображение действительности и которая отличается от первой и своей лексикой[7], и своей семантикой[8], и своим синтаксисом[9], и всем своим эмоциональным тоном. Проявляется ли различие между ними в звфонии, или, иными словами, включаются ли автором в число средств, создающих этот различный тон, и средства фонетические, т.е. звучание? Чтобы ответ на этот трудный вопрос не был необоснованным и субъективным, мы считаем необходимым привести чрезвычайно сухие цифры, характеризующие фонетическую структуру каждой из этих частей, сопоставляя их друг с другом и со всем стихотворением в целом.

Соотношение гласных и согласных в "Незнакомке", как нам известно, почти такое же, как в обычной разговорно-литературной речи. Однако при общем проценте согласных в "Незнакомке" 4I,4, процент их в первой части несколько ниже – 40,2, а во второй соответственно выше – 42,4. Это означает иными словами, что в среднем на строфу при одном и том же количестве гласных (36) согласных приходится в I-й части –53,5, а во второй – 48,9, т.е. на 4,6 меньше. Иными словами, в первой части на один гласный звук приходится I,5 согласных,

а во второй - I,3 согласных. Сама по себе эта разница невелика, но все же говорит о бо́льшей звучности второй части. Это первое указание подтверждается целым рядом других. Так, если сопоставить гласные ударяемые с неударяемыми (а количество слогов в каждой строфе одинаково), то и здесь картина не вполне тождественна: в первой части стихотворения на одну строфу приходится I2 ударений, а во второй - I3. А так как ударяемый слог всегда сильнее и дольше неударяемого, то следовательно и вся вторая часть и в этом отношении звучит иначе, чем первая.

Наконец, если сопоставить гласные по их регистру, то и здесь обнаружатся различия между обеими частями. В среднем на одну строфу приходится:

	Высокие	Средние	Низкие	Итого
I-я часть	I7,2	9,7	9,I	36
2-я часть	I6,7	II,3	8,0	36

Если звуки высокого подъема, узкие, в обеих частях занимают почти одно и то же место, то низких, широких, во второй части меньше, а средних больше. Итак, гласные во второй части сильнее, дольше, и звучат они в среднем регистре, т.е. ровнее, чем в первой.

Неодинаков и состав согласных: мы уже говорили, что во второй части их вообще меньше, чем в первой. Но вторая часть не только поэтому менее шумна, чем первая. В то время как в первой части согласных шумных 67 %, во второй их только 64,6 %, то есть она звучнее. А кроме того, согласных глухих приходится в среднем на одну строфу в первой части 20,I, а во второй - только I7,7. Как ни малы эти цифры сами по себе, но в пределах строфы в сочетании с количеством ударяемых гласных и сонорных согласных это явственно воспринимается слухом как иное звучание.

Итак, по всем показателям в т о р а я ч а с т ь з в у ч н е е п е р в о й, обладает большей слышимостью, силой и звучит в более высоком тоне, т.е. звучание ее соответствует в какой-то степени содержанию.

I0.

В ином направлении, но не менее значительно проявляется в "Незнакомке" организация рифм. Всякая рифма по самой своей природе является упорядочением звуковым и композиционным, организуя повторением тождественных или чрезвычайно близких звуков и строку и строфу. В "Незнакомке" эта орга-

низующая роль рифмы расширена: рифма выходит за пределы одной строфы и распространяется на несколько строф, причем строф не смежных, а находящихся на более или менее далеком расстоянии друг от друга. И этим достигается известная нам музыкальность, которая и здесь проявляется в том, это единство не переходит в тождество или единообразие. Вдобавок эта повторяемость звучания осуществляется на разных словах, так что при сходстве звуковом сохраняется различие смысловое. Через все стихотворение проходит как бы несколько голосов, сходясь и расходясь, напоминая музыкальное голосоведение, хотя и на ограниченном материале.

В качестве примеров можно привести перекличку строф первой и восьмой: над ресторанами - пьяными (прилаг.) мех пьяными (сущ.) - туманами или строф четвертой и одиннадцатой: уключины - приученный поручены - излучины. И вряд ли нечаянностью является, что в обоих этих случаях рифмы повторяются через семь строф, деля все стихотворение на равные части.

К этому второму повторяющемуся созвучию (у́чьнъ) примыкают как бы звуковые варианты его в строфе 2 -переулочной - булочной[+] и в строфе 7 - назначенный - схваченный, отличающиеся от основной рифмы или изменением опорного гласного звука (А взамен У в строфе 7) или перестановкой звуков (строфа 2).

Таким образом, перекликающимися дактилическими рифмами в нечетных стихах охвачены строфы 2, 4, 7 и II, а кроме того I и 8. Но этой перекличкой звучаний связываются между собой обе части стихотворения: ведь половина этих строф (I-2-4) лежит в первой части, а половина (7-8-II) - во второй. Конечно, развитие сюжета-содержания никак из звуковых повторений не вытекает и этих повторений не требует, но наличие повторений подкрепляет тематическую связь еще и связью эвфонической.

Во второй части стихотворения повторяются и мужские

[+] Не исключена возможность, что Блок произносил булошный - переулошный. Такое произношение указано в словаре Ушакова для обоих этих слов. Поскольку, однако, слово уличный произносится только со звуком Ч, а Блок к тому же был не москвич, а петербуржец, мы считаем более вероятным, что и слово переулочный (а, значит, и булочной) он произносил со звуком Ч.

рифмы четных строк, проявляясь в одних случаях как полные созвучия, а в других как диссонансы. Мужскими диссонирующими окончаниями связаны между собой строфы 7-8-13: мне-окне-окна-вручено-вино-мне-вине, где при одном и том же опорном согласном Н (с неизбежным перед Е смягчением) варьируют гласные Е-А-О-Е. При этом в соседних строфах при разных согласных в конечном ударном слоге проходит один и тот же гласный А: одна-окна-мелка-рука-вуаль-даль.

Можно, наконец, привести случай, правда, единственный, где отчетливо ощущается явная и непосредственная связь звучания и значения. Строка 35 из строфы 9 - И шляпа с траурными перьями - и строки 45-46 из строфы 12 - И перья страуса склоненные в моем качаются мозгу связаны между собой этой двойной фонетико-семантической связью. Внешнее звуковое повторение - это повтороение звуков СТРАУ(рными) - СТРАУ(са). Но это звучание раскрывает семантику прилагательного траурный. Мы уже говорили[10], что оно означает не только "погребальный", "печальный", но и "черный". Звуки СТРАУ в строке 35 как бы подготавливают слово страус в строке 45, тем самым подсказывая, что перья эти - страусовые, черные. И в то же время то обстаятельство, что траурный означает "печальный", влечет за собой то, что перья страуса оказываются склоненными, ибо одним из значений слова склоненный является "согбенный", причем не обязательно в физическом плане, т.е. в значении "согнувшийся", а в значении, передающем душевное состояние - удрученный, т.е. печальный, траурный. Так звуковое соответствие создает единый семантический, а в еще большей мере - эмоциональный тон.

Необходимо, однако, отметить, что такое соотношение звуков и смыслов нигде больше в стихотворении не проявляется.

II.

Связывать звучание с содержанием, т.е.предметным, лексическим значением слов, конечно же, нельзя[11], ибо, как не раз мы говорили, это неминуемо грозит субъективизмом и вульгаризацией. Но связать их с "настроением", эмоциональным тоном, по всей видимости, можно, так как в какой-то степени и звучание речи создает это настроение. Но вместе с тем следует иметь в виду, что поэтом выбираются и отбираются не звуки, - выбираются слова, сочетания слов, подбираются повторения звуков. Не исключена возможность, что и самый выбор слов определяется не только их непосредственным

значением, но и звучанием[12], которое, не меная смыслу, уточняет настроение, создает определенный эмоциональный тон. В анализе "Незнакомки" мы пытались показать, как ее звуковая организация - звучность, приглушенность, высокое или низкое звучание, закрытость или открытость звуков - участвует в создании общего тона стихотворения.

Примечания публикатора

1. Здесь и далее для обозначения звуков в статье А.М. Финкеля используются заглавные литеры русского алфавита. Применяемая автором фонетическая транскрипция по условиям задачи совпадает с использованной А.М. Пешковским (её "Образец..." - транскрипцию "I-й тысячи" звуков см. в: Пешковский, 1925, с. 191).

2. Из стихотворения "Чеши томленья", входившего в книгу "Под северным небом", см., напр.: Бальмонт, 1980, с. 40-41.

3. В оригинале первая строка третьей строфы инструментована на [S] : And the silken, sad uncertain rustling of each purple curtain. А.М. Финкель цитирует русский перевод В.Е. Жаботинского, см.: Altalena, 1905, с. 176.

4. См.: Финкель, 1973, с. 320, примечание 3; Финкель, 1961.

5. Данное замечание А.М. Финкеля безусловно несправедливо. Наличие или отсутствие повтора в некотором месте речевой цепи, конечно же, зависит только от того, сколько раз употреблен некоторый звук именно на данном участке. Факторы частоты и употребительности повторяемого звука в данном тексте или в некоторых других, привлекаемых для сравнения, текстах, равно как и факторы сходства или различия повторяемого звука с окружающими его звуками неизбежно должны быть приняты по внимание при определении степени заметности повтора или его функциональной нагрузки в структуре текста. Но О.М. Брик в "Звуковых повторах" решал, главным образом, задачу общей систематизации повторов по их внутренней структуре (гл. 2) и по их локализации ("расстановке") в стиховых единицах (гл. 3). Конкретные примеры, при всем их изобилии, привлекались Бриком не для обсуждения самих этих примеров, а лишь для иллюстрации общей схемы. Поэтому примеры не содержат никаких имплицитных утверждений о заметности или, тем более, особой значимости зафиксированных повторов в тех исходных текстах, из которых эти повторы были извлечены.

6. См. Финкель, 1973, с. 314, 320 и след. Ср ниже в 9 и

нами примечания к нему.

7. Во второй части стихотворения слов "явно сниженных или обиденных... мы уже более не встречаем. Вместо этого на фоне общеупотребительной лексики выделяются слова и словосочетания более высокого тона, подчас повторяющие... поэтическую традицию" (Финкель, 1973, с. 321).

8. Если для первой части "Незнакомки" типично "словоупотребление - прямое, точное, четкое", то "семантико-эмоциональный тон второй части... создают... слова многозначные, употребленные в нескольких прямых и непрямых значениях сразу, причем сплошь и рядом не в узуальных, а окказиональных" (там же, с. 322).

9. Для первой части "Незнакомки" характерна фраза "четкая, краткая, прозрачная по составу, в которой границы синтаксические совпадают со строфическими". Во второй части фраза становится "широкой, осложненной деепричастными и причастными оборотами, охватывающей подчас (вопреки употреблениям автором знакам препинания) несколько строф" (там же, с. 320).

10. См. Финкель, 1973, с. 323.

11. Это высказывание сегодня уже представляется неоправданно категоричным. См. постановку в современной общей лингвистике вопроса об особой "звукоизобразительной системе языка" (Воронин, 1982; там же представительная библиография проблемы), а применительно к звуковой организации русского стиха - крайне соблазнительные по результатам, но еще не прошедшие убедительной проверки исследования А.П. Журавлева (1974).

12. О том, что определенный звуковой комплекс может быть и первым зародышем всего будущего стихотворения, и "поисковым инструментом" в творческом процессе поэта, писал В.Т. Шаламов (1976). О других концепциях "первоядра" стихотворения и процесса работы над ним см.: Гиндин, 1976, с. 149-150 (еще один вариант представлений о "первоядре" - как о "зрительном ряде", "картине" - описан недавно в Боцу, 1982, с. 185-186).

Л И Т Е Р А Т У Р А

Бальмонт К.Д. Избранное. - М., Худ. лит., 1980, - 742 с.

Боцу П. Удивление перед чудом жизни. Беседа с А.Гордовским. - Вопросы литературы, 1982, № 3, с. 177-197.

Брик О.М. Звуковые повторы. (Анализ звуковой структуры стиха). - В кн.: Поэтика. Пг., 18-я Гос. типография, 1919, с. 58-98.

Воронин С.В. Основы фоносемантики. - Л.: Изд. Ленингр. ун-та, 1982. - 244 с.

Гиндин С.И. Послесловие к статье В.Т. Шаламова. - Семиотика и информатика, М., 1976, вып. 7, с. 142-152.

Журавлев А.П. Содержательность фонетической формы поэтического текста. - Вопросы стилистики, Саратов, 1974, вып. 8, с. 41-60.

Пешковский А.М. Десять тысяч звуков. (Опыт звуковой характеристики русского языка как основы для звфонических исследований). - В кн.: Пешковский А.М. Сборник статей: Методика родного языка, лингвистика, стилистика, поэтика. Л.-М., ГИЗ, 1925, с. 167-191.

Федоров А.В. Звуковая форма стихотворного перевода. (Вопросы метрики и фонетики). - Поэтика: Временник Отдела словесных искусств. Л., 1928, вып. [4] с. 45-69.

Финкель А.М. Стилистическое использование служебных слов (на материале стихотворения А. Блока "Незнакомка"). - Межвузовская конференция по исторической лексикологии, лексикографии и языку писателя. Тезисы докл., Л.,1961.

Финкель А.М. Опыт лингвистического анализа стихотворения А. Блока "Незнакомка". Подготовка текста и публ. С.И. Гиндина. - Сборник научных трудов Московского педагогического института иностранных языков, 1973, вып. 73, с. 314-326.

Шаламов В.Т. Звуковой повтор - поиск смысла. - Семиотика и информатика. М., 1976, с. 128-145.

Altalena. Ворон. Поэма Эдгара По. - В кн.: Чтец-декламатор.: Худож. сб. Т. 2. Киев, Типогр. П. Барского, 1905, с. 175-180.

THE SOUND PATTERN OF "AN UNKNOWN LADY" BY A. BLOK

Alexander M. Finkel (1899-1968)

S u m m a r y

The paper continues the author's "An essay in linguistic analysis of "An unknown lady" published in: "Sbornik nauchnych trudov Moskovskogo pedagogičeskogo instituta inostrannych jazykov", M., 1973, v. 73.

This new paper aims at analyzing the phonic composition of "An unknown lady" (Neznakomka) by Alexander Blok, an acknowledged masterpiece of Russian poetry. The principal method used by the author is a comparison of the 'sound count' of the poem in question with the 'sound count' of Russian colloquial speech compiled by Alexej M. Peshkovsky, a great Russian linguist of the 1920s. Though the poem is shown to be a little bit more sonorous; no significant difference between two counts taken as a whole can be traced. But a thorough analysis reveals some local deviations and sound clusters that can have a definite structural and aesthetic value.

The paper is accompanied with a special c o m m e n t a r y by Dr. Sergej Gindin. He tries to mark some insufficiencies in the author's argumentation and to trace the further development of the problems discussed in the paper.

РАБОТА А.М. ФИНКЕЛЯ В ИДЕЙНОЙ ИСТОРИИ ОТЕЧЕСТВЕННЫХ СТАТИСТИЧЕСКИХ ИССЛЕДОВАНИЙ ПОЭТИЧЕСКОЙ РЕЧИ

(Проза как нейтральный "языковой" эталон в метрике и фонике)

С.И. Гиндин

В научном наследии профессора Харьковского университета Александра Моисеевича Финкеля (1899-1968)[+] важное место занимают выполненные им на рубеже 1950-1960-х годов три "опыта" лингвистического анализа стихотворений А. Блока и Э. Багрицкого[++]. Когда в начале 70-х годов я готовил сокращенную публикацию самого раннего из "опытов" (Финкель, 1973), в моем распоряжении имелся лишь один источник текста - машинопись статьи, завершенной, судя по авторской датировке, в 1960 г. Но позднее мне стало известно, что в 1967 г. А.М. Финкель объединил все три "опыта" в небольшую книгу, получившую заглавие "О поэтическом языке". Познакомившись, благодаря любезности Анны Павловны Финкель, с машинописью названной книги, я увидел, что глава о "Незнакомке" пополнена в ней, по сравнению со статьей 1960 г., еще одним, пятым разделом[+++] «Эвфоника "Незнакомки"». Находка эта позволяет, как кажется, заполнить один весьма заметный пробел в истории идей и методов отечественной лингвистической и литературоведческой статистики (что и дало основание предложить «Эвфонику "Незнакомки"» вниманию читателей "Трудов по

[+] Биографические сведения об А.М. Финкеле и анализ его идей в области стилистики и теории перевода см. в прекрасной статье Айзенштока, 1970. Перечень основных собственно-лингвистических работ А.М. Финкеля см.: Булахов, 1978.

[++] О значении этих "опытов" для семантики и типологии видов речи см.: Гиндин (в печати).

[+++] Структура статьи 1960 г. описана мною во вступительной заметке к публикации Финкель, 1973. Новый раздел, надо полагать, был создан или непосредственно в ходе составления книги 1967 г., или незадолго до того. В пользу такой датировки говорит и тот факт, что единственная в наследии Финкеля, помимо «Эвфоники "Незнакомки"», попытка применения количественных методов - пионерское сравнение переводов одного текста (Финкель, в печати) - была предпринята им именно в 1966 г.

лингвостатистике"[+]).

Идея изучения статистических свойств стихотворной речи посредством сравнения со статистическими свойствами прозы, выступающей в качестве своего рода нейтрального эталона, представителя "собственно-языковых" тенденций, была выдвинута применительно к фонетике стиха (Пешковский, 1925) почти одновременно с получением первых убедительных доказательств ее плодотворности для исследования метрики (см.: Шенгели, 1923)[++]. Существенным методическим новшеством явилось при этом то, что А.М. Пешковский, в отличие от своих современников-стиховедов, выбрал в качестве эталона свойства не письменной художественной или деловой прозы, а устной обиходно-разговорной речи (правда, сознательно просеиваемой исследователем при записи через "орфоэпическое сито", см. Пешковский, 1925, с. 170-171).

Не ограничиваясь формулировкой принципов нового подхода, Пешковский создал основу для его практического применения, составив по материалам своих полевых записей частотной список звуков русской разговорной речи, до сих пор не поте-

---

[+] Источником публикации служит упомянутая машинопись книги "О поэтическом языке". При подготовке текста ссылки на литературу уточнены и приведены в форму, принятую в "Ученых записках" Тартуского университета. Сделанные А.М. Финкелем отсылки к другим частям исследования о "Незнакомке" оставлены без изменения, необходимая ввиду раздельной публикации частей расшифровка таких отсылок дается в моем комментарии, помещенном непосредственно вслед за публикацией. Там же содержатся необходимые уточнения и дополнения к отдельным утверждениям автора. Отсылки в комментарию в тексте делаются с помощью арабских цифр, а отсылки к авторским подстрочным примечаниям - с помощью астерисков, как то принято в "Ученых записках" ТГУ.

[++] Необходимость проекции статистических данных о стихе на некоторую эталонную картину возможностей, допускаемых языком, была осознана Б.В. Томашевским в начале 1910-х годов (см.: Томашевский, 1971). Он же впервые реализовал этот подход с помощью построенной им теоретико-вероятностной модели размера (Томашевский, 1918). Однако источником данных для построения эталонной модели ему служила не проза, как у Шенгели, а сам исследуемый поэтический текст.

рявший своего научного и методического значения[+]. Казалось бы, налицо были все предпосылки для того, чтобы новый метод "эвфонических исследований" нашел широкое применение.

Однако этого не произошло. Если в метрике и ритмике сравнение с моделями, рассчитанными на основании данных о прозе, стало одним из самых популярных и действенных способов изучения русского (обзор по состоянию на начало 70-х гг. см.: Гаспаров, 1974), а затем и целого ряда других национальных стихосложений, то в исследованиях по фонике призыв А.М. Пешковского так и не был подхвачен. Подобное положение тем более парадоксально, что в целом для работ по фонике русского стиха в XX в. характерно едва ли не большее разнообразие применяемых методик (см. их беглый обзор в: Гиндин, 1976), нежели для работ в русской метрике.

В чем причины возникновения подобного историко-научного парадокса? Думается, что основную роль тут могли сыграть два фактора: бóльшая эмпиричность процедуры сравнения и бóльшая трудоемкость и монотонность обработки эмпирического материала.

В самом деле, в метрике сопоставления эмпирических частот различных форм размера проводится обычно не прямо с частотами тех же форм в прозе, а с рассчитанными на основе данных о прозе теоретико-вероятностными "языковыми" моделями размера[++]. "Речевые модели" размеров, основанные непосредственно на статистике употребления стихоподобных отрезков в прозе (см.: Холшевников, 1973), гораздо менее популярны и используются в основном в методических целях для верификации языковых моделей. Основной операцией при расчете теоретической модели является вычисление вероятности

[+] Сведения о других публикациях, содержащих полные или частичные частотные списки русских звуков и звукосочетаний см.: Панов, 1967, с. 81; Ермоленко, 1970, с.53-58; Никонов, 1966, с. 287-288. Отметим, что в обширной прикнижной библиографии в Панов, 1967, статья А.М. Пешковского отнесена (посредством графического выделения) к числу наиболее важных работ по русской фонетике, появившихся за два с лишним века изучения последней.

[++] Наиболее тщательное описание процедуры расчета см.: Červenka, Sgallová, 1967, а также Гиндин, 1982, § I.2-I.3, где введены некоторые коррективы, связанные с возможной неоднородностью метрического материала.

"ритмической вариации" (или "модуляции") размера по данным о частотах употребления в прозе составляющих этой вариации - "ритмических типов слов".

Но как раз эта операция оказывается невозможной в фонике - ведь звук, в отличие от единиц метрики, не имеет синтагматической организации. Поэтому исследователь фоники, изучая статистику звуков, принужден пользоваться аналогом именно "речевых моделей" - сравнивать эмпирию стиха с эмпирией прозы. Правда, теоретико-вероятностную модель можно было бы рассчитать для единиц следующего уровня - звукосочетаний. Но здесь она из-за большого числа единиц и многообразия позиционных ограничений будет существенно уступать моделям размеров в обозримости и удобстве использования.

В то же время первичная обработка материала, получение сравниваемых эмпирических данных в фонике более трудоемки и требуют бо́льших специальных знаний. Исследователь русской метрики должен уметь отличать ударный слог от безударного. Исследователь фоники должен давать полную квалификацию отдельным конкретным звукам. Многие методы анализа фоники эту трудность обходят, предлагая ограничить анализ лишь некоторыми типами звуков (чаще всего гласными). В отличие от них подход Пешковского, нацеленный на создание полной статистической картины звукоупотребления, предполагает сплошное транскрибирование текста, а это занятие большинству наших филологов, не являющихся по своей узкой специальности фонетистами, и поныне представляется чуждым и скучным.

То, что подход, предложенный Пешковским, оказался реализован именно А.М. Финкелем, подтверждает нашу гипотезу: как раз для него оба отмеченных препятствия были несущественны. Применение количественных методов в его работах вырастало именно из эмпирического материала и ограничивалось простейшими операциями, диктуемыми самим материалом, вопросы же теоретического и методологического осмысления этих операций с позиций математики перед ним, по-видимому, даже не возникали. В то же время, хотя он и не был фонетистом, для него как русиста универсального профиля не составляло никаких трудностей обработка фонетического материала.

Что же дал этот долгожданный опыт реализации идей Пешковского?

Основные результаты публикуемой работы, как представляется, сводятся к убедительной демонстрации двух следующих

фактов:

1. Фонетический состав поэтического произведения в целом может и не отличаться от средне-статистических норм обиходной речи (ср., однако, бо́льшую звучность поэзии, отмечавшуюся и самим Пешковским).

2. Даже при отсутствии подобных отличий внутри поэтического произведения возможны композиционно и экспрессивно значимые локальные перегруппировки звуков. В обнаружении таких перегруппировок А.М. Финкель опирается опять-таки на статистику, фактически используя те же принципы, которые были примерно в те же годы испробованы в анализе ритмики А.М. Колмогоровым и А.М. Кондратовым (1962; о внутренних ограничениях "девиационного" подхода см. в: Гиндин, 1970).

Конечно, сегодня, спустя 20 лет после написания публикуемой работы, в ней легко увидеть и серьезные методические изъяны. С точки зрения статистики это в первую очередь отсутствие оценок значимости констатируемых А.М. Финкелем различий в фонетическом составе между частями или строфами стихотворения. С точки зрения поэтики это ощутимая изоляция анализа фонетической композиции стихотворения от анализа других композиционных планов, противостоящая преобладающим сегодня тенденциям к "комплексному", или "целостному", анализу поэтического текста. Впрочем, последний недостаток, возникший, конечно, из-за описанного временного разрыва в написании работы, был бы в значительной степени ослаблен, доведись исследованию А.М. Финкеля сразу увидеть свет в полном составе.

Л И Т Е Р А Т У Р А

Айзенштох И.Я. А.М. Финкель - теоретик художественного перевода. - В кн.: Мастерство перевода, 1970. М.: Сов. писатель, 1970, с. 91-118.

Булахов М.Г. Финкель Александр Моисеевич. - В кн.: Булахов М.Г. Восточнославянские языковеды. Т. 3. Минск: Изд. Белорус. ун-та, 1978, с. 252-254.

Гаспаров М.Л. Квантитативные методы в русском стиховедении. - В кн.: Гаспаров М.Л. Современный русский язык: Метрика и ритмика. М.: Наука, 1974. с. 18-38.

Гиндин С.И. Пути моделирования ритмической организации текста. - В кн.: Структурно-математические методы моделирования языка. Ч. I, Киев, 1970, с. 33-35.

Гиндин С.И. Послесловие к статье В.Т. Шаламова. - Семиотика и информатика. М., 1976, вып. 7, с. 147-152.

Гиндин С.И. Ритмика, интонация и смысловая композиция в поэме Вл. Луговского "Как человек плыл с Одиссеем". - В кн.: Проблемы структурной лингвистики, 1978. М.: Наука, 1982, с. 230-265.

Гиндин С.И. "Опыты лингвистического анализа" А.М. Финкеля и проблема соотношения словарных и текстовых значений в поэзии и прозе. (В печати).
Ермоленко Г.В. Лингвистическая статистика: Краткий очерк и библиогр. указатель. Алма-Ата, 1970. – 156 с.
Колмогоров А.Н., Кондратов А.М. Ритмика поэм Маяковского. – Вопросы языкознания, 1962, № 3, с. 62-74.
Никонов В.А. Фоностатистическое измерение междуязыковых расстояний. – В кн.: Исследования по фонологии. М.: Наука, 1966, с. 285-297.
Панов М.В. Русская фонетика. – М.: Просвещение, 1967. – 438 с.
Пешковский А.М. Десять тысяч звуков. (Опыт звуковой характеристики русского языка как основы для звфонических исследований). – В кн.: Пешковский А.М. Сборник статей: Методика родного языка, лингвистика, стилистика, поэтика. – Л.-М.: ГИЗ, 1925, с. 167-191.
Томашевский Б.В. [Письма В.Я. Брюсову 1910-1911 гг.]. Вступит. статья и публикация Л.С. Флейшмана. – Ученые записки Тарт. ун-та, 1971, вып. 284. Труды по знаковым системам, 5, с. 532-544.
Томашевский Б.В. Ритмика 4-стопного ямба по наблюдениям над стихом "Евгения Онегина". – Пушкин и его современники, Пг., 1918, вып. 29-30, с. 144-187.
Финкель А.М. Опыт лингвистического анализа стихотворения А. Блока "Незнакомка". / Подготовка текста и публ. С.И. Гиндина. – Сборник научных трудов Московского педагогического института иностранных языков, 1973, вып. 73, с. 314-326.
Финкель А.М. "Ночная песнь странника" Гете в русских переводах. / Публ. и вступит. статья С.И. Гиндина. (В печати.)
Холшевников В.Е. Случайные четырехстопные ямбы в русской прозе. – In: Slavic Poetics. The Hague – P.: Mouton, 1973, pp. 549-557.
Шенгели Г.А. Трактат о русском стихе. Ч. I, Изд. 2-е, перераб. М.-Пг.: ГИЗ, 1923. – 184 с.
Červenka M., Sgallová K. On a probabilistic model of the Chech verse. – Prague studies in mathematical linguistics, 1967, [v.] 2, pp. 105-120.

PROFESSOR FINKEL'S PAPER IN THE HISTORY OF IDEAS OF RUSSIAN LITERARY STATISTICS OF PROSE AS A NEUTRAL LINGUISTIC STANDARD IN METRICAL AND PHONETIC STUDIES OF RUSSIAN VERSE

Sergej I. Gindin

S u m m a r y

This paper reveals a certain paradox in the development of ideas and methods of Russian literary statistics. The idea that the statistical structure of verse can be studied (on different levels) by comparing it to that of prose was put forward in metrics and phonetics almost simultaneously. But while in metrical studies this idea has become a widely used instrument of studying various versification systems, in phonetics of verse it hasn't received any practical application.

The author tries to find the roots of the paradox, and argues that Prof. Finkel's paper (also published in this volume) can be treated as a first attempt to fill the gap.

СОДЕРЖАНИЕ

Из истории лингвостатистики:

RESUMEES - SUMMARIES

Ученые записки Тартуского государственного университета.
Выпуск 711.
КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ 1985.
На русском языке.
Резюме на немецком и английском языках.
Тартуский государственный университет.
ЭССР, 202400, г.Тарту, ул.Ülikooli, 18.
Ответственный редактор Й. Тулдава.
Подписано к печати 26.07.1985.
МВ 07076.
Формат 60x90/16.
Бумага писчая.
Машинопись. Ротапринт.
Учетно-издательских листов 9,15. Печатных листов 9,5.
Тираж 500.
Заказ № 716.
Цена 1 руб. 40 коп.
Типография ТГУ, ЭССР, 202400, г.Тарту, ул.Пийбеова, 17.
6 - 2