ISSN 0494-7304 0207-4435

TARTU ÜLIKOOLI

TOIMETISED

УЧЕНЫЕ ЗАПИСКИ ТАРТУСКОГО УНИВЕРСИТЕТА

ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS

912

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

1990

QUANTITATIVE LINGUISTICS AND AUTOMATIC TEXT ANALYSIS

TARTU 1990

TARTU ÜLIKOOLI TOIMETISED
УЧЕНЫЕ ЗАПИСКИ ТАРТУСКОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS
Alustatud 1893.a. VIHIK 912 ВЫПУСК Основаны в 1893.г.

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

1990

QUANTITATIVE LINGUISTICS AND AUTOMATIC TEXT ANALYSIS

TARTU 1990

Kogumik "Kvantitatiivlingvistika ja tekstide automaatanalüüs" ilmub Tartu Ülikooli rakendus- ja arvutuslingvistika uurimisgrupi iga-aastase väljaandena alates 1985.a. (jätkates sarja "Töid keelestatistika alalt" I-X, mis ilmus 1976-1984). Käesolevas kuuendas (16-ndas) väljaandes (1990) on avaldatud kõrgkoolide vahelise probleemrühma "Tekst interdistsiplinaarse uurimise objektina" liikmete artiklid.

Сборник "Квантитативная лингвистика и автоматический анализ текстов" публикуется Группой прикладной и компьютерной лингвистики Тартуского университета начиная с 1985 г. (сборник является продолжением серии "Труды по лингвостатистике" I-X, 1976-1984 гг.). Настоящий 6-й (16-й) выпуск (1990 г.) содержит статьи членов Межвузовской проблемной группы "Текст как объект междисциплинарных исследований".

The collections "Quantitative Linguistics and Automatic Text Analysis" appears annually since 1985, edited by the members of the Applied and Computational Linguistics Group at Tartu University (Estonia). The collections continue the series "Papers on Linguo-Statistics" I-X (1976-1984). The present issue No. 6 (16) contains investigations by members of the All-Union Research Group "Text as an object of interdisciplinary investigations".

Ученые записки Тартуского университета.
Выпуск 912.
КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ
АНАЛИЗ ТЕКСТОВ 1990.
QUANTITATIVE LINGUISTICS AND
AUTOMATIC TEXT ANALYSIS.
На русском языке.
Резюме на разных языках.
Тартуский университет.
ЭР, 202400, г.Тарту, ул.Юликооли, 18.
Ответственный редактор Ю. Тулдава.
Подписано к печати 6.XI.1990.
Формат 60x90/16.
Бумага писчая.
Машинопись. Ротапринт.
Учетно-издательских листов 7,63. Печатных листов 7,75.
Тираж 500.
Заказ № 745.
Цена 2 руб. 30 коп.
Типография ТУ, ЭР, 202400, г.Тарту, ул.Тийги, 78.
6 - 2

КВАНТИТАТИВНОЕ ИССЛЕДОВАНИЕ ПОЛИСЕМИИ КОРНЕВЫХ СЛОВ РУССКОГО ЯЗЫКА XI-XX ВЕКОВ

А.В. Андреевская

Данное исследование базируется на машинной версии Словаря корневых слов русского языка XI-XX веков (далее - СКС), созданного автором в рамках программы формирования Машинного фонда русского языка. Этот словарь представляет собой многоаспектное описание русских корневых слов, то есть слов, представляющих собой, с точки зрения их морфемного состава, свободные корни, обычно осложненные флексией (в том числе и нулевой). Каждой единице (их 5 858) в словаре приписан стандартный набор признаков-параметров, охватывающих фонетические, грамматические, семантические, словообразовательные характеристики корневых слов, а также - хронологию их существования, происхождение и употребительность (всего 7 зон). Каждый из этих параметров включает в себя и более дробные признаки. Так, в семантической зоне содержится, помимо собственно толкований, сведения о числе значений слова и код семантического поля. При анализе данных машинного словаря корневых слов признак исследуется как с содержательной стороны, так и квантитативно, анализируется взаимосвязь параметров.

Ниже речь пойдет лишь об одном параметре из семантической зоны - о числе значений корневого слова.

Центром исследования полисемии корневых слов является определение общих - синхронных и диахронических - закономерностей. Материалом для анализа служат данные, полученные машинной обработкой полной выборки корневых слов.

Так как СКС охватывает широкий временной интервал, вся совокупность единиц описания стратифицирована по пяти периодам: 1 период - древнерусский, XI-XIУ вв.; 2 - среднерусский, XУ-XУП вв.; 3 - XУШ в.; 4 - XIX в.; 5 - XX в. Такая периодизация принята и при описании семантики корневых слов.

В таблице I показано распределение корневых слов по числу значений в каждый из пяти рассматриваемых периодов.

Достаточно сложно определить, какой закон наиболее точно описывает распределение N(m), так как от 92% (в первый период) до 98% (в пятый) слов имеют не более трех значений. Для всех периодов кроме первого доля слов N с числом значений m составляет от 0,5 до 1,0 % при любом $m > 5$. При таком распределении выбор функции p(m), аппроксимирующей N(m) по крите-

Таблица I

Распределение корневых слов по числу значений (в абсолютных цифрах и в % к общему числу корневых слов в данный период)

число значений	период 1		2		3		4		5	
	абс.	%	абс.	%	абс.	%	абс.	%	абс.	%
I	I682	70.30	232I	78.60	28I4	82.28	3385	82.36	3397	8I.I7
2	343	I4.30	428	I4.50	456	I3.45	529	I2.87	586	I4.00
3	I69	7.06	I27	4.30	90	2.63	I30	3.I6	I4I	3.37
4	82	3.4I	39	I.32	35	I.02	39	0.96	37	0.88
5	45	I.90	2I	0.7I	II	0.32	I5	0.36	II	0.26
6	25	I.04	9	0.30	8	0.23	2	0.05	5	0.I2
7	I8	0.75	2	0.07	3	0.09	6	0.I5	2	0.05
8	6	0.25	3	0.I0	I	0.03	0	0.00	4	0.I0
9	9	0.38	I	0.03	I	0.03	3	0.07	I	0.02
I0	2	0.08	0	0.00	I	0.00	0	0.00	0	0.00
II	5	0.2I	I	0.03			I	0.02	0	0.00
I2	2	0.08	0	0.00					I	0.02
I3	2	0.08	I	0.03						
I4	I	0.04								
I5	0	0.00								
I6	I	0.04								
I7	0	0.00								
I8	I	0.04								
I9	0	0.00								
20	I	0.04								
Всего слов	2304		2953		3420		4II0		4I85	

рию абсолютной близости (например, по минимуму суммы квадратов абсолютных отклонений $\Sigma (N(m)-p(m))^2$), неоднозначен. Такое же положение наблюдается и для русского языка в целом (см. таблицу 2).

Таблица 2

Распределение корневых слов по числу значений в пятый период в сопоставлении с распределением для всего современного русского языка в целом (по выборке из "Словаря русского языка" С.И. Ожегова) в %

число значений	корневые слова	современный русский язык		
		в целом	существительные	
			в целом	бесприставочные
I	8I.I7	73.I8	77.70	80.00
2	I4.00	I8.8I	I6.8I	I5.66
3	3.37	4.68	3.64	2.98
4	0.88	I.9I	0.98	0.83
5	0.26	0.75	0.46	0.33
6	0.I2	0.45	0.23	0.08
7	0.05	0.07	0.I2	0.08
8	0.I0	0.05	0.06	
9	0.02	0.05		
I2	0.02	0.00		
I5		0.02		

Разными исследователями предлагаются разные формулировки закона распределения доли слов в зависимости от числа значений: Ю.К. Крылов считает, что эта зависимость либо логарифмическая (Крылов Ю.К., Якубовская М.Д., I977), либо описывается дискретным вариантом нормального закона (Крылов Ю.К., I982); Ю.А. Тулдава предлагает в качестве универсальной формулу $p(m)=ae^{-b\sqrt{m}}$, где m - число значений, p(m) - доля слов с таким числом значений, a и b - коэффициенты (Тулдава Ю.А., I987).

Переход к минимизации относительных отклонений $(\Sigma(\{N(m)-p(m)\}/p(m))^2)$ позволяет подобрать аппроксимацию p(m), близкую к N(m) во всем диапазоне изменения m. Расчеты показывают, что наибольшую близость по относительному отклонению (с относительной дисперсией 0.05-0.08 для m от I до I2) обеспечивает выбор функции распределения p(m) вида $p(m)=ae^{-b\sqrt{m}}$ при a=I5÷25 и b =3÷3.5 в зависимости от периода (наименьшие значения коэффициентов относятся к первому периоду, далее они постепенно возрастают). Этот закон соответствует выведенному Ю.А. Тулдава для современного русского и некоторых других языков (для русского языка a=II.4, b =2.9). Такое совпадение, с одной стороны, подтверждает универсальность данной закономер-

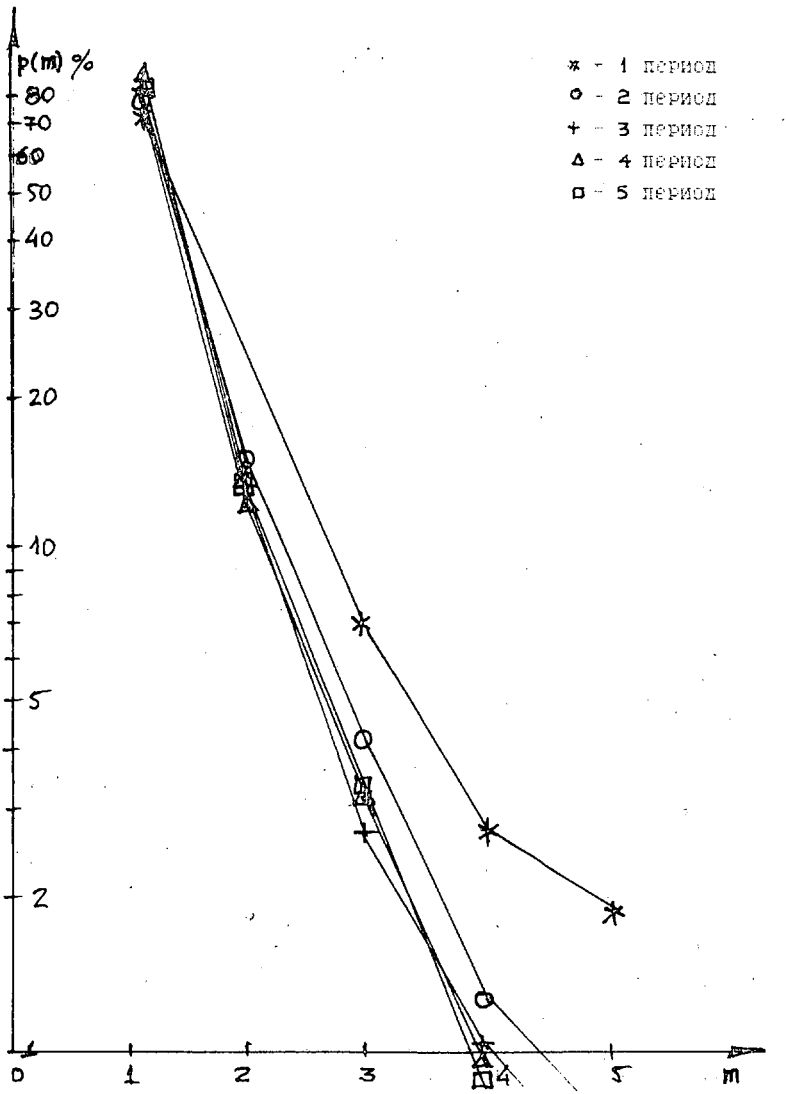

Рис. 1. Распределение корневых слов $p(m)$ по числу значений m (для $m < 5$) по периодам

ности, а с другой, показывает, что с точки зрения полисемии корневые слова представляют собой подсистему языка, своего рода проекцию всей лексики, только в измененном масштабе.

Масштаб этот задается двумя разными характеристиками – одной формальной, связанной с принципами описания значений в СКС, второй – качественной, вытекающей из свойств самого рассматриваемого объекта.

А.А. Поликарповым и О.С. Крюковой отмечено, что количество слов с заданной полисемией зависит от типа словаря(Поликарпов А.А., Крюкова О.С., 1989). Словарь корневых слов по составу своего словника представляет собой выборку из большого словаря и, следовательно, параметры левой части словарной статьи не должны приводить к уменьшению доли многозначных слов. Но зона семантики в СКС не является прямым отображением правой части "Словаря современного русского литературного языка". Это вызвано тем, что при формировании лексикографического описания корневых слов возникла необходимость соединить в одной словарной статье толкования из словарей разных эпох и разных авторов. Так как для русского языка XVШ и XIX веков пока не существует полных научных описаний лексики, пришлось обратиться к старым словарям, созданным на основе иных принципов. Несмотря на известные недостатки таких описаний (например, в них не фиксируются оттенки значений), именно на этой основе пришлось унифицировать все толкования корневых слов в СКС. В результате в этом аспекте он оказался близок к современным кратким словарям типа "Словаря русского языка" С.И. Ожегова (в табл. 2 и 3 приведены цифры, полученные на основе данных А. А. Поликарпова и О. С. Крюковой (Поликарпов А.А., Крюкова О.С., 1989)).

Описанный принцип унификации описания значений корневых слов, вполне приемлемый для XУ-XX веков, оказался трудно применимым к толкованиям древнерусских слов. Для языка этого периода не существует критерия, позволяющего определить меру сводимости лексико-семантических вариантов в одно словарное толкование. Более того, не всегда можно отличить окказиональные словоупотребления от типичных. В результате древнерусские корневые слова кажутся более многозначными. Это обстоятельство необходимо учитывать при квантитативном анализе полисемии корневых слов.

Второй фактор, влияющий на распределение корневых слов по числу значений связан непосредственно с составом описывае-

мого множества языковых объектов, а именно с преобладанием в нем имен существительных (они составляют от общего числа корневых слов от 81 % первый период до 95 % в пятый). Общеизвестно, что степень полисемичности слов разных частей речи различна, в частности, у существительных она ниже, чем у глаголов. Верно это и для корневых слов. Таким образом, преобладание существительных приводит к сдвигу соотношений между группами слов с разным числом значений в сторону увеличения доли наименее многозначных слов.

Оба отмеченных обстоятельства сказались и на средних значениях индекса многозначности корневых слов.

Таблица 3

Средние значения индекса многозначности для корневых слов по периодам и для современного русского языка (по выборке из "Словаря русского языка" С.И. Ожегова)

корневые слова					современный русский язык		
период					в целом	существительные	
I	2	3	4	5		в целом	бесприставочные
I.7	I.3	I.25	I.25	I.26	I.4I	I.3I	I.26

Приведенные в таблице 3 данные показывают, что индекс многозначности является для корневых слов практически постоянной величиной, неизменной с XУ века. Отличие ее значений для первого периода от всех остальных связано с уже упоминавшейся выше невозможностью полностью унифицировать толкования древнерусских слов в соответствии с общими принципами описания значений в СКС.

При сопоставлении средней многозначности корневых слов с соответствующими величинами для современного русского языка обращает на себя внимание незначительность расхождения между ними. Особенно примечательно абсолютное совпадение этих цифр для корневых слов и для бесприставочных существительных.

Таким образом, анализ распределения корневых слов по числу значений в сопоставлении с данными по русскому языку в целом доказывает, что каждая подсистема языка, в том числе и корневые слова, подчиняется общим законам такого распределения, но при этом испытывает на себе и влияние таких факторов как частеречная принадлежность слова и сложность его морфемного состава.

Особый интерес представляет изучение исторической динамики семантического объема корневых слов. Эти данные позволяют уточнить некоторые из ранее сделанных выводов.

В таблице 4 приведены значения рассматриваемого признака для пар соседних периодов. Признак принимает следующие значения: "=" (число значений у слова при его переходе из периода T в период T+I осталось неизменным). ">" (число значений увеличилось) и "<" (число значений уменьшилось).

Таблица 4

Историческая динамика параметра "число значений" (для пар соседних периодов)

	пары периодов			
	I → 2	2 → 3	3 → 4	4 → 5
=	II23 (70 %)	I733 (84 %)	2673 (89 %)	3279 (86 %)
>	I0I (7 %)	I04 (6 %)	I98 (7 %)	242 (6.5%)
<	372 (23 %)	226 (I0 %)	II4 (4 %)	I58 (4 %)

Из таблицы 4 видно, что среди корневых слов очень высок процент слов с неизменным числом значений (от 70 до 89 %).

Аналогичное положение наблюдается и при анализе полного цикла жизни этих слов: 69 % корневых слов, существующих более двух периодов, на протяжении всей своей истории не изменили числа значений, у 0.4 % семантический объем увеличился и всего у 0.03 % – уменьшился. (У остальных 30 % корневых слов общей тенденции в изменении числа значений не выявлено).

Данные таблицы 4 указывают на то, что при переходе из первого периода во второй корневые слова активно "теряют" свои значения. В дальнейшем этот процесс постепенно замедляется и к XУШ веку стабилизируется окончательно. Это явление тесно связано с отмеченной выше разницей индексов полисемичности слов первого и последующих периодов, с несколько иной конфигурацией кривой распределения древнерусских корневых слов по числу значений. И объяснение этого явления, вероятно, такое же: словари XУ–XУП и XУШ веков фиксируют меньшее число значений, пренебрегая их контекстными вариантами.

В целом же с точки зрения полисемии корневые слова весьма стабильны на протяжении всей своей истории. Единственная заметная граница проходит между первым и вторым периодами и связана она, в первую очередь, с расхождениями в принципах словарных описаний.

Подводя итог этому фрагменту исследования русских корневых слов XI–XX веков, суммируем сделанные наблюдения:

I. Было отмечено, что распределение корневых слов в зависимости от числа значений наиболее точно аппроксимируется универсальным законом, выведенным Ю.А. Тулдава: $p(m) = ae^{-b\sqrt{m}}$, где m – число значений, a=I5÷25, b =3÷3.5 в зависимости от периода. Именно этот закон позволяет учитывать особенности распределения корневых слов p(m) при m>5.

2. Среднее число значений корневых слов является характеристикой, стабильной для XУ–XX веков (его значения, соответственно: I.3, I.25, I.25, I.26). Древнерусские слова имеют более высокое значение индекса многозначности (I.7), что связано с расхождениями в типах словарного описания древнерусского языка и языка более поздних эпох.

3. Не только индекс полисемичности всего множества корневых слов, но и число значений каждого отдельного слова стремится к стабильности: от 70 до 89 % корневых слов имеют неизменный семантический объем.

При переходе из первого периода во второй число значений сократилось у 23 % рассматриваемых лексем. Это явление тесно связано с уменьшением в тот же период индекса многозначности и имеет те же причины.

Сходство данных по корневым словам и по современному русскому языку в целом позволяет предполагать, что многие из выводов, полученных на материале корневых слов, можно(с учетом влияния морфемного строения слова на его семантический объем) отнести ко всему русскому языку.

Л И Т Е Р А Т У Р А

Крылов Ю.К., Якубовская М.Д. Статистический анализ полисемии как языковой универсалии и проблема семантического тождества слова // НТИ. Сер. 2. – I977. – № 3. – С. I–6.

Крылов Ю.К. Об одной парадигме лингвистических распределений // Учен. зап. Тартуского гос. университета. Вып. 268. – Тарту, I982. – С. 80–I02.

Поликарпов А.А., Крюкова О.С. О системном соотношении краткого и среднего толковых словарей русского языка // Учен. зап. Тартуского гос. университета. Вып. 872. – Тарту, I989. – С. III–I25.

Тулдава Ю.А. Проблемы и методы квантитативно-системного исследования лексики. – Тарту, I987.

RUSSIAN XI-XX CENTURIES ROOT-WORDS QUANTITATIVE ANALYSIS

Alina W. Andreewskaya

S u m m a r y

It is the semantic volume of 5 858 Russian root-words that is analysed. The dependence of their distribution law on the number of their meanings is stated. Changes of average polysemy in the period from XI to XX centuery and of separate root-words semantic volume is outlined too.

The results show that root-word's semantic volume is a stabile characteristic. Author states that analysed set of words is regulated by same laws as total vocabulary.

МЕТОДЫ АВТОМАТИЧЕСКОЙ АТРИБУЦИИ ДОКУМЕНТОВ: ПРАКТИЧЕСКИЕ РЕЗУЛЬТАТЫ

М.С. Блехман

При построении локальных проблемно-ориентированных систем автоматизированной обработки научно-технической информации (ЛАСНТИ) нередко возникает необходимость определения тематики поступающих в систему документов. Это вызвано неполным соответствием структуры информационной потребности организации-пользователя ЛАСНТИ общесоюзному рубрикатору ГАСНТИ. В этом случае некоторая рубрика локального рубрикатора может частично пересекаться с несколькими рубриками ГАСНТИ. Иными словами, документы, относящиеся к такой рубрике, рассеяны по нескольким рубрикам общесоюзного рубрикатора, причем последние содержат не только документы, релевантные информационной потребности организации, но и шумовые. Проиллюстрируем сказанное примером Всесоюзного научно-исследовательского и опытно-конструкторского института технологии электромашиностроения - "ВНИИТэлектромаш" (г. Харьков).

Таблица I

Соотношение рубрик ГАСНТИ и ЛАСНТИ "Технология" института "ВНИИТэлектромаш"

Рубрики ЛАСНТИ / Рубрики ГАСНТИ	Обмоточно-изолировочные процессы	Пропиточно-сушильные процессы	Сборочные процессы	Контрольные процессы	Производство коллекторов	Конструкция электродвигателей и их узлов
I	2	3	4	5	6	7
45.01 - Общие вопросы электротехники	+	-		+	-	-
45.09 - Электротехнические материалы	+	+	-	-	-	-
45.29 - Электрические машины	+	+	+	+	+	+
45.31 - Электрические аппараты	+	+	-	+	-	-
45.33 - Трансформаторы	+	+	+	+	-	-
45.47 - Провода и кабели	+	-	-	-	-	-

В этой таблице задано соответствие рубрик верхнего уровня локального рубрикатора рубрикам второго, среднего уровня ГАСНТИ.

При построении АСНТИ в организации, рубрики информационной потребности которой неполностью соответствуют рубрикам ГАСНТИ, можно рекомендовать использование процедуры автоматической атрибуции (ААД) документов. Целью внедрения ААД является:

- отсев шумовых документов,
- присвоение релевантным документам индексов рубрик ГАСНТИ и, как результат, упрощение формулирования поисковых предписаний.

Создание системы ААД требует от ее разработчиков компромисса между качеством работы системы и возможностью ее тиражирования. Опыт разработки различных локальных АСНТИ позволяет автору утверждать, что оптимальным подходом к решению этой задачи является разработка программного обеспечения, относительно легко, без участия лингвиста настраиваемого на соответствующую предметную область. Зависимость же системы от профессионального лингвиста существенно затрудняет ее тиражирование, хотя и может в некоторых случаях дать весьма высокое качество атрибуции (Пиотровский, Шингарева и др. 1985).

Харьковский творческий коллектив разработки информационных систем, включающий специалистов ВНИИТэлектромаша ВНИИОМШСа (Всесоюзный НИИ организации и механизации шахтного строительства) и научно-исследовательской группы вычислительной лингвистики Харьковского университета, а также Института повышения квалификации информационных работников (ИПКИР, г. Москва) разрабатывает тиражируемые системы обработки текстовой информации. Эти системы строятся методом "выращивания", т.е. постепенного приближения к оптимальному балансу эффективности и тиражируемости.

Разрабатываемая с 1982 г. система ААД прошла несколько этапов развития.

1. С 1982 г. по 1988 г. путем "выращивания" была создана система атрибуции[+] основанная на методе дизъюнкции (Певзнер, Блехман, Аксельрод). Функционирование этой системы основано на предложенном Б.Р. Певзнером методе эталонных тематических массивов.

[+] Программисты - Е.М. Бахарева, А.Е. Аксельрод, Д.Б.Кушнарева

Эталонный массив документов формируется следующим образом.

I) На вход системы поступают магнитная лента, сорержащая рефераты документов в формате МЕКОФ, и перечень релевантных тематик - рубрик верхнего уровня локального рубрикатора.

2) Программным путем из каждого документа выделяется его заголовок либо (по требованию пользователя) заголовок и реферат. На печать выводятся выделенные тексты, и при каждом из них распечатываются: (а) порядковый номер документа; (б) перечень номеров релевантных тематик (пример приведен на рис. I).

ОБРАБАТЫВАЕМЫЙ ДОКУМЕНТ № 367

УСТРОЙСТВО ДЛЯ ОСАДКИ ОБМОТКИ И ОПРЕССОВКИ ИЗОЛЯЦИИ В СЕРДЕЧНИКЕ ЭЛЕКТРИЧЕСКОЙ МАШИНЫ

I 2 3 4 5 6 7 8 9 I0 II I2 I3 I4 I5
I6 I7 I8 I9 20 2I 22 23 24 25 26 27 28 29 30

ОБРАБАТЫВАЕМЫЙ ДОКУМЕНТ № 368

УСТРОЙСТВО ДЛЯ ФОРМОВАНИЯ ЛОБОВЫХ ЧАСТЕЙ ОБМОТКИ

I 2 3 4 5 6 7 8 9 I0 II I2 I3 I4 I5
I6 I7 I8 I9 20 2I 22 23 24 25 26 27 28 29 30

Рис. I. Фрагмент эталонного массива документов.

3) Эксперт обводит номер тематики, к которой относится документ, после чего в систему вводится массив номеров документов с соответствующими номерами тематик.

4) Программным путем из исходного массива формируется массив документов, каждому из которых присвоено имя тематики.

Из эталонного массива программным путем формируется частотный словарь основ. Для учета того факта, что тематики содержат различное количество документов, подсчитывается относительная частота термина - отношение абсолютной частоты к количеству документов данной тематики.

5) В заголовке и реферате каждого документа выделяются словоформы; от них отсекаются окончания.

6) Слова сортируются по алфавиту; повторы отбрасываются и подсчитывается количество документов, в которых встретилась данная основа.

7) Формируется выходной массив - частотный словарь для

классифицирования документов дизъюнктивным методом. Приведем фрагмент словаря (рис. 2).

КАПЕЛЬН			5				
КАТУШЕК	6			2			
КАТУШК	4						
КАЧЕСТВ	I			I0			
КЛИНЬ	3						
КОЛЛЕКТОР	2				32		
КОЛЛЕКТОРН				2	9	I	
КОММУТАЦ				4	3		
КОМПАУНД			20				
КОНТАКТ					3		
КОНТАКТН					3		
КОНТРОЛ		I	5	23	3		I
КОНТРОЛЬН				2			
КОРОТК		I		2			
КОРОТКОЗАМКНУТ	I	I				I	
КОРПУС		I		2		I	
КОСВЕННН				2			
КРЕПЛЕН	3	2					

Рис. 2. Фрагмент частотного тематического словаря.

Объем словаря в ЛАСНТИ "Технология" - около I тыс. основ. Общеупотребительные и окказиональные основы удалены из сформулированного системой словаря лингвистом. Опыт показывает, что такая лингвистическая работа является необходимой, т.к. наличие в словаре общеупотребительных слов существенно влияет на качество атрибуции.

Атрибуция документов осуществляется следующим образом.

На вход системы поступает магнитная лента с документами в формате МЕКОФ. Из очередного документа выделяются задаваемые пользователем текстовые поля: заголовок или заголовок и реферат. В текстовых поля отыскиваются основы, зафиксированные словарем. Если таковые отсутствуют, документ отсеивается в "шум". Если же словарные основы найдены, то подсчитывается их суммарный "вес" по каждой из рубрик локального рубрикатора. Документ относится к той рубрике, по которой он набрал наибольший вес. Приведем пример. Имеем документ

КОНТРОЛЬ КАЧЕСТВА КОЛЛЕКТОРОВ ЭЛЕКТРИЧЕСКИХ МАШИН

Из этого документа в словаре (рис. I) нашлись следующие основы:

Рубрика ЛАСНТИ Основа	Намотка	Сборка	Пропитка	Контроль	Производство коллекторов	Двигатели	Шум
КАЧЕСТВ	I			IO			
КОЛЛЕКТОР	2				32		
КОНТРОЛ		I	5	23	3		I
ИТОГО	3	I	5	33	35		I

Поскольку максимальный вес набрали документом по рубрике "Производство коллекторов", система относит его к этой рубрике и вносит ее индекс в документ.

Описанная технология ААД реализована пакетом прикладных программ АСТРИ ("Автоматизированная система тематического распределения информации"), для ЭВМ типа ЕС. Полный цикл обработки одного документа требует 3-4 сек. процессорного времени ЭВМ ЕС-I035 при анализе заголовка и реферата и менее I сек. - при анализе только заголовка. В течение I988 г. пакетом АСТРИ было обработано не менее 20 тыс. документов по электротехнике. Средняя точность атрибуции - около 85%. Потери релевантных документов минимальны.

В то же время анализ работы пакета позволил выявить ряд недостатков реализованного в нем алгоритма.

а) Словарь системы состоит из изолированных слов, их контекст не учитывается. Результатом этого являются ошибки при отнесении документов к тематикам.

б) В словарь системы попадает очень большое количество случайных, в том числе общеупотребительных, слов. Их наличие в словаре недопустимо, т.к. дает непредсказуемые ошибки при отнесении документов к тематикам, поэтому при создании словаря необходимо вручную удалять эти слова.

в) В словаре системы задаются веса слов. Опыт показал, что весовые характеристики также нуждаются в ручной корректировке.

г) Наконец, реализованный алгоритм позволяет относить документ только к одной тематике, тогда как не менее IO% документов относятся более чем к одно рубрике верхнего уровня локального рубрикатора. Так, документ "Контроль качества коллекторов электрических машин" относится к двум рубрикам "Производство коллекторов" и "Контроль".

Был сделан вывод о том, что реализованный в АСТРИ алгоритм ограничивает возможности тиражирования системы, делая необходимым авторский надзор на этапе создания словаря.

При этом работа со словарем требует большого объема квалифицированной ручной работы при практически гарантированном количестве ошибок на этапе автоматического определения тематики документов (I5-20%).

2. Для устранения указанных недостатков в I989 г. начата разработка новой версии АСТРИ. В этом варианте система использует модификацию предложенного Б.Р. Певзнером метода конъюнкции. Словарь системы в новой редакции должен состоять из цепочек основ слов с приписанными каждой цепочке номерами тематик, к которым эта цепочка относится. Кроме того, существует список основ, встретившихся в цепочках. Оба словаря не являются частотными, т.е. ни основы, ни их цепочки не содержат весовых характеристик.

Формирование словаря цепочек начинается с формирования списка основ.

а) Сначала пользователю распечатываются документы обучающего массива (в том виде, как это реализовано в АСТРИ). Пользователь отбирает релевантные документы, не производя их классифицирования.

Примечание: Если заранее известно, что в массиве не может быть шумовых документов, то работа начинается непосредственно со следующего этапа (см. этап (б)).

б) Из отмеченных документов система формирует список основ слов без окончаний. Веса не приписываются. Отбрасывание слов по весовому порогу не производится. Допускается, но не обязательно, ручное отбрасывание ошибочных и случайных слов.

в) Затем из заданного пользователем поля каждого релевантного документа автоматически формируется список всех возможных цепочек.

Примечание. В тексте реферата цепочки формируются внутри предложений.

- Заранее вручную заготавливаются модельные комбинации цепочек для всех типов текстов. Тип текста-количество в нем основ из списка, сформированного на предшествующих этапах.

Пример.

Имеем текст типа "4", т.е. в нем содержится 4 основы из списка. Модельная комбинация для данного типа текста имеет следующий вид:

I234, I23, I2, 234, 23, 34, I34, I3

Если бы имели текст типа "3", то имели бы такую модельную комбинацию:

I23 I2 23 I3

Примечание I. Здесь каждая цифра (I,2,3,4 и т.д.) означает, что некоторая основа является в данном тексте первой (второй, третьей, четвертой) из найденных в списке.

Примечание 2. Для упрощения дальнейшей работы системы в качестве параметра задается максимальное значение типа текста (по умолчанию - 5). Тогда исходный текст автоматически разбивается на цепочки по 5 основ, найденных в списке.

- В соответствующую модельную комбинацию подставляются порядковые номера основ.

2) Полученные таким образом цепочки основ нумеруются системой и выводятся на печать вместо с "меню" номеров тематики, аналогично выводу документов.

На печать каждая цепочка выводится в виде:

ЦЕПОЧКА №
ОСНОВА (ЕЕ ПОРЯДКОВЫЙ НОМЕР В СПИСКЕ ОСНОВ)
НОМЕРА ТЕМАТИК

Пример:

ЦЕПОЧКА № I
УСТРОЙСТВ (288) НАТЯЖЕН (I95) ПРОВОД (207)
I 2 3 4 5 6 7 8 . . . . . .

д) Специалист анализирует каждую цепочку основ так, как если бы это был заголовок документа, и обводит номер соответствующей тематики (тематик). Количество тематик, к которым может относиться одна цепочка, не должно превышать 3. Если цепочка не может быть отнесена к какой-либо одной, двум или трем тематикам, то ни один из номеров не обводится.

е) После того как специалист обработал всю распечатку, данные его анализа вводятся в систему в виде:

№ цепочки № тематики, № тематики

Система формирует на МД словарь цепочек в виде:

Цепочки типа I:
Цепочка Тематика
Цепочки типа 2:
Цепочка Тематика I Тематика 2
Цепочки типа 3:
Цепочка Тематика I Тематика 2 Тематика 3

Здесь цепочка - это последовательность порядковых номеров основ в списке основ. Например, для приведенной выше цепочки

устройств (288) натяжен (I95) провод (207)

имеем на диске цепочку:

288 I95 207 I

где I - номер некоторой тематики ("Обмоточно-изолировочные процессы").

ж) В завершение формирования словаря система исключает из списка основ те основы, которые не встретились ни в одной из цепочек. При этом автоматически корректируются цепочки, т.к. изменились порядковые номера основ.

Автоматическая атрибуция осуществляется путем выявления в заданном пользователем поле документа (например, заголовке и / или таких- то предложениях текста) цепочек, содержащихся в словаре. Считается, что документ относится к той тематике (тематикам), к которой (которым) относятся найденные цепочки. Если же ни одной такой цепочки в документе не найдено, то он отсеивается как шумовой. Алгоритм классифицирования имеет следующий вид:

I) Поиск в документе основ из списка. Если при поиске по заголовку найдено не более одной основы, то либо отсеиваем документ как шумовой, либо ищем основы в реферате, анализируя либо каждое предложение, либо первые I0 слов текста.

2) Определение типа документа в зависимости от количества найденных в списке основ. Порождение всех возможных цепочек.

3) Поиск цепочек в словаре на полное совпадение. Приоритет более длинных цепочек, т.е. если в словаре нашлась, например, 4-элементная цепочка, то 3- и 2-элементные цепочки не ищутся.

4) Приписывание документу тематики (тематик) найденных цепочек.

ЛИТЕРАТУРА

Певзнер Б.Р., Блехман М.С., Аксельрод А.Е. Методы распределения документов по тематическим базам данных без использования рубрикаторов // НТИ, сер. I, 1987, № 6. - С. I0-I5.

Пиотровский Р.Г., Шингарева Е.А., Серебряков А.А., Белякова И.П., Смольникова Е.К. О лингвистическом аппарате машинного анализа единиц и связей текста // Инженерная лингвистика и романо-германское языкознание: Межвузовский сборник научных трудов. - Л.: ЛГПИ им. А.И. Герцена. I985. - С. 98-II2.

SOME METHODS OF AUTOMATIC TEXT ATTRIBUTION:
PRACTICAL RESULTS

Mikhail S. Blekhman

S u m m a r y

Some problems are discussed of working out and implementing automatic topic recognition routines in documentary systems. Methods of computer-based dictionary formation are outlined which provide compilation of frequency dictionaries of isolated word stems as well as contextual non-frequency ones. Algorythms of topic attribution are described.

СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ КОММУНИКАТИВНЫХ СВОЙСТВ ВОПРОСОВ И ОТВЕТОВ РУССКОЙ ДИАЛОГИЧЕСКОЙ РЕЧИ

Н.И. Голубева-Монаткина

I. Свойства вопросов и ответов диалогической речи неоднократно исследовались в различных аспектах и с помощью различных методов. В результате получены различные характеристики этих свойств, в том числе характеристики лексико-синтаксические, логико-семантические (семантико-логические), логические, трансформационные, информационные, науковедческие, социолингвистические. Далее излагаются и лингвистически интерпретируются некоторые статистические характеристики коммуникативных свойств вопросов и ответов, т.е. таких сторон вопросов и ответов, которые, обуславливая различия и общность вопроса с вопросом, ответа с ответом и обнаруживаясь лишь в отношении вопроса с вопросом, ответа с ответом, формируются в соответствии с той или иной установкой говорящего в процессе коммуникации.

Были построены две матрицы: матрица вопросов, столбцы которой образованы свойствами вопросов, а строки - самими вопросами, и матрица ответов со столбцами, формируемыми свойствами ответов, и строками, образованными самими ответами. Матрицы обработаны на ЭВМ ЕС-1022, ЕС-1035 пакетом прикладных программ СОМИ. С помощью расчета средних арифметических, стандартного отклонения, а также корреляционного анализа определялись характеристики частотности проявления и меры вариации результатов по данному свойству в отобранных массивах вопросов и ответов, связи каждого свойства с каждым другим и связи каждых вопроса и ответа с каждым другим.

II. Статистические характеристики свойств вопросов. Расчет средних арифметических (М) показал, что самым распространенным является собственно-вопрос - таким свойством могут обладать 78% вопросительных предложений (М = 0,783). Закономерность высокой частотности проявления этого свойства обусловлена тем, что именно собственно вопрос отражает основную коммуникативную установку спрашивающего и тем самым воплощает сущность вопросительного предложения - его направленность на выяснение неизвестного компонента ситуации. Достаточно часто спрашивают и в том случае, когда одинаково неизвестны несколько компонентов ситуации (Каковы должны быть мой пер-

вые слова? Товарищи? Граждане? Друзья?)[+] Об этом свидетельствует высокая проявляемость в анализируемом массиве свойств вопроса, в котором все неизвестные одинаково неизвестны (M = 0,531). Примерно 44% вопросов непосредственно касается того, к кому обращаются, являются ты-/вы-вопросами (M = = 0,438).

Довольно редко спрашивающим выясняется факт наличия или отсутствия действия, состояния, признака (Павел уехал? 'уехал или не уехал'[++])(M = 0,354). Немногим более четверти вопросов могут констатировать данное положение вещей, передавать предположение, догадку (Ты не обижаешься на меня, так ведь?) (M = 0,288). Совсем редко спрашивающий задает вопрос, в котором одно из неизвестных ему представляется наиболее вероятным (Это кто же сказал? Толик?), вопросы сопоставляющий (А на такси?) и уподобляющий (И на такси не успеете?). Эти свойства характеризуются величинами средних арифметических в пределах от M = 0,38 до M = 0,162.

Расчет стандартного отклонения (s), позволяющего проанализировать вариативность каждого свойства в отобранном массиве, или, иначе, однородность - разнородность массива вопросов в отношении каждого свойства, выявил, что наибольшей величиной стандартного отклонения обладают вопросы, в которых все неизвестные компоненты ситуации одинаково неизвестны (s = 0,5). Именно это свойство больше всего варьирует в данном массиве, различая наибольшее количество рассматриваемых вопросов. Достаточно высокую величину s имеют ты-/-вы-вопросы (s = 0,497), вопросы, в которых неизвестен факт наличия-отсутствия действия, состояния, признака, (s = = 0,479) и вопросы-констатации данного положения вещей (s = = 0,454).

Несколько меньшие значения стандартного отклонения - у вопроса, в котором одно из неизвестных наиболее вероятно, сопоставляющего и уподобляющего вопроса, вопроса-контактной формы (Спрашивающий: Все пройдет, все забудется - слышите? Отвечающий: Вы думаете?), вопроса-констатации противоположного положения вещей (Кто себе зла желает? 'никто себе зла не желает') (s = менее 0,4). Лингвистическое объяснение такого значения дисперсии состоит в многозначности лексико-

---

[+] Здесь и далее источники примеров не указываются.

[++] На выделенном слове находится т.н. логическое ударение.

грамматического состава предложений, которые могут иметь перечисленные свойства (Он им позвонит? а) 'позвонит или не позвонит', б) 'он или не он', 'им или не им', в) 'он им позвонит, так ведь'?, г) в очень специальном контексте 'пусть он им позвонит'), и их способности вступать в синонимические отношения (Павел приедет туда завтра. А вы? = А вы тоже туда придете завтра? = И вы тоже туда приедете завтра? = И вы туда приедете завтра? = И вы?).

Корреляционный анализ матрицы вопросов позволил проверить гипотезу о тех коммуникативных свойствах вопросов, которые можно считать существенными: наличие связей свойства с другими свидетельствует о том, что данное свойство является существенным для вопросов, отсутствие же корреляций с другими говорит о несущественности данного свойства. Существенными в данном массиве признаются свойства с величиной коэффициента корреляции (r) не ниже 0,25. Таких свойств оказалось девять.

I) Собственно-вопрос связан положительно с вопросами, в которых все неизвестные компоненты ситуации одинаково неизвестны, (r = 0,406) и отрицательно с вопросами, констатирующими данное или противоположное положение вещей (r = -0,344; r = -0,382). Это обусловлено тем, что вопрос, в котором все неизвестные компоненты ситуации одинаково неизвестны, всегда является собственно-вопросом (Где же это ты пропадаешь?). Такой вопрос, констатирующий данное и противоположное положение вещей, используются говорящим в противоположных ситуациях: первые - для выявления неизвестного в ситуации, вторые - для констатации уже известной ситуации. Собственно-вопрос положительно связан с вопросом, в котором неизвестен факт наличия или отсутствия действия, состояния, признака, (r = = 0,282) - последний сам является собственно-вопросом. Отрицательная корреляция собственно-вопросов с вопросами-контактными формами (r = -0,207) объясняется тем, что вопросы-контактные формы редко используются как собственно-вопрос.

2) Вопрос, в котором все неизвестные одинаково неизвестны, положительно связан с собственно-вопросом (r =0,406), поскольку всегда им является. Коэффициент корреляции, однако, не слишком высок - некоторые типы предложений, передающих вопрос, в котором все неизвестные одинаково неизвестны, имеют многозначный лексико-грамматический состав и могут использоваться для передачи несобственно-вопросов.

Анализируемое свойство связано отрицательно с вопросом-констатацией данного положения вещей – их зоны распространения не пересекаются, так как в первом случае вопрос задается о неизвестном компоненте ситуации, а во втором в вопросительной форме констатируется все данная ситуация (τ = -0,439): Когда Таня звонила? Утром, днем или вечером? и Разве Таня звонила утром? Данное свойство связано отрицательно с вопросом, в котором неизвестен факт наличия или отсутствия действия, состояния, признака (τ = - 0,384). Вопросы с такими свойствами используются в разных ситуациях.

3) Вопрос, в котором одно из неизвестных наиболее вероятно, связан отрицательно с вопросом, в котором все неизвестные компоненты ситуации одинаково неизвестны (τ = -0,301) – спрашивающим эти вопросы обычно используются разных ситуациях (ср. Куда Таня идет? и Куда Таня идет? В театр?). Можно отметить слабую положительную связь данного вопроса с вопросом-констатацией данного положения вещей (τ = 0,178). Она обусловлена тем, что из всех типов собственно-вопросов именно вопрос, в котором одно из неизвестных наиболее вероятно, имеет несколько сниженную степень неизвестности, а это сближает его с таким несобственно-вопросом, как вопрос-констатация данного положения вещей.

4) Вопрос, в котором неизвестен факт наличия или отсутствия действия, состояния, признака, связан отрицательно с вопросами с вопросительным словом (τ = -0,599) и вопросами, в которых все неизвестные компоненты ситуации одинаково неизвестны (τ = -0,384). Это обусловлено тем, что вопрос, в котором неизвестен факт наличия или отсутствия действия, состояния, признака, не может передаваться предложением с вопросительным словом (но лишь предложениями типа Павел уехал? Павел уехал или не уехал?, Павел уехал или нет?), а последнее характерно именно для вопроса, в котором все неизвестные одинаково неизвестны.

Характерна слабая связь данного свойства с собственно-вопросом (τ = 0,282). Это объясняется многозначностью лексико-грамматического состава предложения типа Павел уехал?, наиболее, по-видимому, распространенного при передаче этого вопроса из трех только что перечисленных – Павел уехал? может передавать не только собственно вопрос, но и констатацию данного положения вещей. По этой же причине рассматриваемое свойство имеет слабую корреляцию и с вопросом, в котором одно из неизвестных наиболее вероятно (τ = 0,205).

5) Сопоставляющий вопрос имеет в качестве обязательного компонента средств выражения вопросительную частицу а (А Павел?). Это отражено в его положительной связи с вопросами с вопросительными частицами (r = 0,538). Сопоставляющий вопрос положительно коррелирует с условными вопросами (r = 0,266), поскольку вопросы с если часто бывают сопоставляющими (А если их нет?). Предложение, передающее сопоставляющий вопрос, практически не способно служить повторением вопроса при ответе. Это подтверждается отрицательной связью свойств сопоставляющего вопроса и повторения при ответе (r = -0,243). Небезынтересна положительная связь рассматриваемого свойства с вопросом, в котором все неизвестные одинаково неизвестны – предложения, имеющие эти свойства, синонимичны (Павел когда придет? и А Павел когда придет?).

6) Уподобляющий вопрос положительно связан с вопросом-констатацией данного положения вещей (r = 0,288), поскольку обладает пониженностью степенью неизвестного. Он как бы находится на границе с несобственно-вопросами, а значит и с вопросом-констатацией данного положения вещей (ср.И ты идешь туда? и Как? (И) Ты идешь туда?!). Отрицательная корреляция уподобляющего вопроса с вопросом, имеющим поясняющее значение (А Павел, он уехал?), (r = -0,244) обусловлена практической невозможностью передать поясняющее значение в предложениях, выражающих уподобляющий вопрос.

7) Вопрос-повторение при ответе положительно коррелирует со свойствами вопросов, имеющих вопросительные слова. Это – вопросы, в которых все неизвестные одинаково неизвестны (r = 0,206), вопросами составными (Где находится город Тимбукту и каково его население?) (r = 0,153) и многоместными (Какие из мальчиков являются братьями каких девочек?) (r = = 0,149). Вопросы с этими свойствами могут служить повторениями при ответе. Но такая возможность характеризует не все вопросительные предложения, имеющие перечисленные свойства – нередко их лексико-грамматический состав препятствует использованию вопроса как повторения при ответе (Ну какая была обувь, например, на Руси?). Именно подобные случаи делают величины соответствующих коэффициентов корреляции низкими.

Свойства вопроса-констатации данного положения вещей, сопоставляющего вопроса, некодифицированности, ты-/вы-вопроса имеют отрицательную корреляцию с рассматриваемым свойством (от r = 0,272 до r = -0,222), поскольку предложения с перечисленными свойствами в качестве повторения при ответе не употре-

8) Предложение, передающее вопрос-констатацию данного положения вещей, не может иметь вопросительное слово (τ = = -0,551) вступать в синонимические отношения с предложениями-собственно-вопросами (τ = -0,344), вопросами, в которых все неизвестные одинаково неизвестны, (τ = - 0,439) и сопоставляющими (τ = -0,182). Предложения, имеющие данное свойство, не используются в качестве вопросов-повторений при ответе (τ = -0,272). В вопросе-констатации данного положения вещей альтернативы задаются перечислением - оба свойства связаны положительно (τ = 0,212). Также положительно данное свойство связано с уподобляющим вопросом (τ = = 0,288) и вопросом, в котором неизвестен факт наличия или отсутствия действия, состояния, признака (τ = 0,178). Последняя корреляция обусловлена тем, что одно и то же предложение может иметь оба эти свойства: Павел приехал? 'Павел приехал, так ведь?' и 'Павел приехал или не приехал'.

9) Вопрос-констатация противоположного положения вещей отрицательно связан с собственно-вопросом (τ = -0,382) - в большинстве случаев одно и то же предложение не может иметь оба эти свойства. Например, Павел приехал? передает собственно-вопрос двух типов (Приехал Павел или не Павел?, Павел приехал или не приехал?) и констатацию данного положения вещей (Павел приехал?!), но не констатацию противоположного положения вещей. В вопросе-констатации противоположного речь идет, в основном, не о том, к кому обращаются, а о 1-м или 3-м лице. Это подтверждается как отрицательной связью данного свойства с ты-/вы-вопросом (τ = -0,185), так и его положительной связью с я-вопросом. Вопросы-констатации противоположного очень редко бывают некодифицированными (τ = = -0,149) - "обыденной" речи т.н. риторические вопросы не свойственны.

Ш. Статистические характеристики свойств ответов. Отвечающий всем другим предпочитает ответ, подтверждающий полностью или частично представление ситуации, отраженное в вопросе (С. Курение вредно или аморально? О. Курение вредно, но не аморально.) (М = 0,445). В исследуемом массиве имеют низкую частотность ответы, которые опровергают представление ситуации, данное спрашивающим (С. Что ты мерзнешь? О. Я я не мерзну.) (М = 0,255). Это может объясняться общей тенденцией отвечающих в большинстве случаев так или иначе соглашаться с представлением ситуации, предлагаемым спрашивающим (ту же тенденцию отмечают социологи при опросах).

Предпочтение, оказываемое подтверждающим ответам, может быть связано с тем, что такие ответы требуют от говорящего меньше психических и интеллектуальных усилий, чем ответы опровергающие.

Сравнительно распространены "свернутые" ответы (наличие-отсутствие этого свойства определялось не по отношению к возможному в позиции ответа повествовательному "полному" предложению, а по отношению к вопросу): (о поликлинике Семашко) С. Взрослая, детская? О. Взрослая. Они составляют 42% (M = = 0,420). Примерно каждый четвертый ответ характеризуется свойством неинформативности (С. И за кого к ты хочешь выйти замуж? О. Не скажу.) (M = 0,25). Этим свойством могут обладать не только "неответы" (см. последний пример), но и собственно-ответы (С. Что ты сказал? О. Я сказал то, что сказал). "Неответом" является каждый пятый из исследуемых (M = = 0,19). Это показывает, что отвечающий не столь уж редко уклоняется от ответа, не имея по каким-то причинам возможности или желания удовлетворить информационную потребность спрашивающего. Отвечающий может давать также и больше информации, чем необходимо спрашивающему - в отобранном массиве доля информативно избыточных ответов составляет 16% (M=0,16). Этим свойством нередко обладают некодифицированные ответы (С. Что/разреза́ть хочешь? О. И разреза́ть/ и класть корицу/ и завертывать/).

Наиболее разнообразен массив ответов по свойствам ответов подтверждающего (s = 0,498) и "свернутого" (s = 0,495), эти свойства в наибольшей степени варьируют среди отобранных ответов, различая наибольшее их количество. Средние значения стандартного отклонения имеют опровергающий ответ (s = 0,419), собственно-ответ (s = 0,431) и неинформативность ответа (s = = 0,434). Более низкими величинами дисперсии, т.е. низкой различающей способностью, обладают "неответ" (s = 0,393) и "вопросительный" ответ (С. Он уже уехал? О. А почему тебя это интересует?) (s = 0,368). Более низки значения стандартного отклонения у остальных свойств ответов (от s = 0,301 до s = 0,218). В целом результаты расчета стандартного отклонения свидетельствуют о большой однородности обсуждаемого массива в отношении этих свойств.

Корреляционный анализ матрицы ответов выявил, что из предполагаемых существенными двадцати свойств ответов действительно существенными являются лишь пятьнадцать.

I) Собственно-ответ связан положительно с ответами подтверждающим и опровергающим (τ = 0,510; τ = 0,279), это - собственно-ответы. Положительная корреляция рассматриваемого свойства ответов с критерием оценки ответов по их информативной избыточности (τ = 0,249) свидетельствует о том, что примерно четверть собственно-ответов информативно избыточна.

2) Подтверждающий ответ - это часто собственно-ответ (τ = 0,510), да-ответ (τ = 0,339). Положительно связан подтверждающий ответ с свойством ситуационной неинформативности ответа (τ = 0,280) - более, чем в одном случае из четырех отвечающий подтверждает имеющееся у спрашивающего представление ситуации, но в данной ситуации ответ не несет необходимой информации. На естественную противопоставленность данного свойства опровергающему и нет-ответу указывают отрицательные коэффициенты корреляции (τ = -0,482; τ = -0,290). Подтверждающий ответ во многом противоположен "неответу" (τ = - 0,382). Очень редко рассматриваемый ответ бывает неинформативным вне зависимости от ситуации (τ = -0,354) либо передается с помощью вопросительного предложения (τ = -0,281).

3) Опровергающий ответ естественным образом связан положительно с нет-ответом (τ = 0,601) и отрицательно с ответом подтверждающим (τ = -0,482). Отвечающие довольно часто опровергают представление ситуации, заданное спрашивающим, с помощью правила логического вывода (τ = 0,371) (С. Ты разбил стекло? О. Стекло разбил Иванов. 'стекло разбил не я').

4) "Выводной" ответ нередко бывает информативно избыточным (τ = 0,475) и опровергает представление ситуации, имеющееся в вопросе (τ = 0,371). Например, С. Где проводятся Олимпиады по языковедению и математике для школьников? О. Последний раз какие бы то ни было олимпиады проводились во П веке до н.э.

5) Ответ-контактная форма (С. Что ты говоришь? И он никуда не уезжал? О. Представляешь?) обладает свойствами "вопросительного" ответа (τ = 0,549) и "неответа" (τ = 0,493). Он бывает нередко неинформативным вне зависимости от ситуации (τ = 0,410). Ответ-контактная форма обычно не может быть собственно-ответом (τ = -0,321) и "свернутым" ответом (τ = -0,224).

6) Свойство повторения вопроса при ответе (С. Вы много убили львов, мсье Тартарен? О. Много ли я их убил?) положительно связано с "вопросительным" ответом (τ = 0,526) и "неответом" (τ = 0,474). "Неответность" повторения вопроса при

ответе подтверждается также его отрицательной корреляцией с собственно-ответом (τ = -0,243) и подтверждающим ответом (τ = -0,205).

7) Свойство не знаю-ответа отрицательно связано с собственно-ответом (τ = -0,403) и подтверждающим ответом (τ = -0,205). Не знаю-ответы обладают свойством "свернутости" (τ = 0,177).

8) Свойство "неответа" положительно связан с показателями внеситуационной неинформативности (τ = 0,839), "вопросительного" ответа (τ = 0,797), контактной формы (τ = 0,493), повторения вопроса при ответе (τ = 0,474) - ответы, имеющие одно из свойств, могут считаться "неответами". "Неответ" отрицательно коррелирует с собственно-ответом (τ =-0,702), подтверждающим ответом (τ = -0,382) (последние не могут квалифицироваться как "неответы"). "Неответам" несвойственно быть "свернутыми" (τ = -0,335), информативно избыточными (τ = -0,211). Да- и нет-ответы не являются "неответами" (τ = -0,161; τ = -0,157).

9) Свойство информативной избыточности имеют ответы, являющиеся собственно-ответами (τ = 0,240): "выводной" (τ = 0,475), опровергающий (τ = 0,255), подтверждающий (τ =0,232). Не пересекаются в отобранном массиве зоны распространения ответов информативно избыточных и внеситуационно неинформативных (τ = -0,252), "неответов" (τ = -0,211).

10) Свойством внеситуационной неинформативности характеризуются "неответы" (τ = 0,839), контактные формы (τ = 0,410), ответы-повторения вопроса при ответе (τ = 0,398) и "вопросительные" (τ = 0,661). Неинформативные вне зависимости от ситуации ответы - это не те ответы, которые охарактеризованы как несобственно-ответы (τ = -0,718), подтверждающие (τ = -0,354) или опровергающие (τ = -0,175). Ответ вряд ли может иметь одновременно свойства внеситуационной неинформативности и "свернутости" (τ = -0,328).

11) Свойством быть неинформативными (С. Кто построил этот дом? О. Самый богатый человек в городе.) обладает часть собственно-ответов (τ = 0,162), в том числе подтверждающий (τ = 0,280). Вопросы ситуационно неинформативные по зоне своего распространения нередко противоположны вопросам неинформативным вне зависимости от ситуации.

12) "Вопросительный" ответ способствует различению ответов, имеющих-не имеющих каждое из следующих четырех свойств - "неответа" (τ = 0,797), внеситуационной неинформативности

(τ = 0,661), контактной формы (τ = 0,549), повторения вопроса при ответе (τ = 0,526). "Вопросительные" ответы, в основном, не бывают собственно-ответами (τ = -0,512), ответами "свернутыми" (τ = -0,344) или подтверждающими (τ = -0,281).

13) Свойство да-ответа естественным образом характеризует ответы подтверждающие (τ = 0,339) и "свернутые" (τ = 0,189). Свойство быть да-ответом во многом противоположно свойствам "неответа" (τ = -0,161), внеситуативной неинформативности (τ = -0,154), "вопросительного" ответа (τ =-0,145), опровергающего ответа (τ = -0,140). Эти свойства обычно не могут принадлежать одному и тому же ответу.

14) Свойство нет-ответа положительно связано с опровергающим ответом (τ = 0,601). Нет-ответы являются "свернутыми" (τ = 0,308), довольно часто собственно-ответами (τ = 0,185). Нет-ответ характеризует ответы, отличающиеся от подтверждающих (τ = -0,290) либо неинформативные вне зависимости от ситуации (τ = -0,187). Нет-ответ - это не "неответ"(τ =-0,157).

15) Свойство "свернутости" может иметь нет-ответ (τ = 0,208), да-ответ (τ = 0,189), не знаю-ответ (τ =0,177). Свернутость" и "вопросительный" ответ обычно характеризуют разные ответы (τ = -0,344) - так же, как "сверкутость" и "неответ" (τ = -0,335), внеситуационной неинформативности (τ = -0,328), контактная форма (τ = -0,224), повторение вопроса при ответе (τ = -0,195).

16. Статистические характеристики коммуникативных свойств вопросов и ответов служат более глубокому познанию этих свойств. Обоснованное Н.Р. Кэмпбеллом измерение свойств объектов с помощью присваивания цифр для представления свойств на нашем материале дало возможность выявить сходства и различия внутри каждой пары свойств, увидеть одно отношение сходства внутри разных пар и разные отношения сходства внутри одной пары.

STATISTICAL CHARACTERISTICS OF COMMUNICATION PROPERTIES OF QUESTIONS AND ANSWERS OF RUSSIAN DIALOGIC SPEECH

N.I. Golubeva-Monatkina

S u m m a r y

The article states and gives a linguistic interpretation of some statistic characteristics of communicative properties of questions and answers. The matrix of questions and answers is processed by the computer. Characteristics of frequency of occurence and measure of variation of results concerning the given property among the selected questions and answers correlations of each property with each other, correlations of each question with each other as well as correlations of each answer with each other are determined with the help of calculating arithmetical means, standard deviation and correlational analysis.

ИЗОМОРФНЫЕ И ОТЛИЧИТЕЛЬНЫЕ ЧЕРТЫ МОРФЕМЫ И СЛОГА В РАСПРЕДЕЛЕНИИ ДЛИНЫ

Е.И. Гороть

Данная статья отражает результаты статистического исследования длины морфемы (корневой, префиксальной и суффиксальной) и слога в фонемах. Данные единицы были вычленены из прилагательных современного английского языка, полученных в результате сплошной выборки из словаря The Random House Dictionary of the English Language (1975).

Длина единиц языка в фонемах - одна из характеристик их фонемной структуры. Предполагается, что длина морфемы и слога зависит от их фонемной структуры и от начальной фонемы.

Для проверки первого предложения исследуемые единицы были записаны в виде канонических форм (КФ),в которых вместо конкретной гласной стоит символ V, а вместо согласной - символ C (см. об этом: Hocket, 1958, p. 284). Анализ инвентаря КФ показал, что длина корневой морфемы колеблется от одной до девяти фонем, суффикса и слога - от одной до шести, и префикса - от одной до пяти фонем. Корневые морфемы описываются 127-ю КФ, префиксы - 18-ю, суффиксы - 17-ю и слоги - 16-ю КФ, каждая из которых имеет определенную частотность. Но, как справедливо подчеркивал Н.С. Трубецкой, "абсолютные цифры фактической частотности имеют лишь второстепенное значение. Реальную значимость имеет только соотношение этих чисел с числами, выражающими теоретически рассчитанную частотность" (Трубецкой, 1960, с. 294).

Теоретически возможные КФ исчислялись по формуле 2^n, где n - количество фонем в исследуемой единице (Перебийніс, 1970, с. 160).

Количество теоретически возможных КФ растет очень быстро с увеличением длины, однако язык не использует всех теоретически возможных КФ (на исследуемом материале процент реализации равняется 29,03% в префиксе, 12,70% в суффиксе и слоге и 12,43% в корне). При этом наблюдается такая закономерность: чем длиннее исследуемая единица,тем меньшая часть от теоретически возможных КФ используется языком. Одни КФ не употребляются потому, что они не могут образовать морфему или слог, поскольку подобные сочетания фонем противоречат законам английского языка. Другие структуры допускаются

языком, но на исследуемом материале не зафиксированы. В суффиксе не встретились структуры, содержащие более, чем две, а в префиксе и слоге – более, чем три согласных подряд. Наибольшая последовательность гласных равняется двум фонемам во всех массивах, кроме слога, естественно. Общими для всех массивов являются структуры V, VC, CV, VCC, CVC, CVCC.

Для выявления структурных особенностей языка важно не только установить инвентарь КФ исследуемых единиц, но и определить роль каждой КФ в языке, т.е. установить частоту каждой разновидности фонемной структуры.

Анализ частоты КФ определенной длины показал,что в каждой совокупности несколько КФ описывают большое количество исследуемых единиц, в то время как остальные КФ – низкочастотны. Так, например, в совокупности двухфонемных единиц самая высокая частота у структуры VC в префиксе и суффиксе и CV в корне и слоге. В трехфонемных единицах во всех массивах, кроме суффикса, самой частотной является модель CVC, а в суффиксе преобладает структура VCC. Среди четырехфонемных единиц наиболее частотными оказались структуры CVCC в корне, CCVC в слоге, CVCV в префиксе, VCVC в суффиксе. Самыми частотными пятифонемными структурами являются CCVCC в префиксе и слоге, CVCVC в корне, VCVCV в суффиксе. Шестифонемные корни отдают предпочтение структурам типа CVCCVC, а в суффиксе только одна шестифонемная структура – CVCVCC. При образовании семифонемных корней превалируют структуры типа CVCVCVC, восьмифонемных – CVCVCCVC, девятифонемных – CVCVCVCVC.

Чем объяснить, что именно эти КФ обладают самой высокой частотой? Очевидно, различный вес КФ в языке объясняется структурными свойствами самих КФ, закономерностями их построения. Так, среди высокочастотных КФ отсутствуют такие, в которых имеются стечения нескольких гласных подряд. Мало КФ со скоплениями согласных, превалируют структуры с равномерным расположением согласных относительно гласных. Если сочетания согласных имеются, то они находятся в начале или в конце структуры. Следовательно, в моделирующей силе КФ большую роль играет наличие и длина степений фонем одного класса, а также позиция, которую занимают гласные и согласные в данных структурах.

На рис. I графически изображено распределение длины исследуемых единиц. Такое изображение дает возможность учесть одновременно несколько характеристик: длину исследуемых еди-

ниц и частоту, характер изменения частоты с изменением длины.

Как видим, все кривые распределения длины имеют асимметрический характер. Наблюдается сдвиг влево, в сторону меньших значений. У всех кривых по одному пику-вершине.

По характеру графика исследуемые единицы делятся на две группы - корень и слог, с одной стороны, префикс и суффикс, с другой.

Кривые распределения длины корня и слога более плавны. Пик слабо выражен. Отличие этих кривых состоит в местонахождении пика: пик кривой корня находится на трехфонемных единицах, в слога - на двухфонемных.

Для кривых распределения длины аффиксальных морфем характерен резкий подъем и спад. Как префикс, так и суффикс имеют пик на двухфонемных единицах. Различие состоит в высоте пика - у префикса высота пика ниже.

По максимальной частоте наблюдается иное распределение: в префиксе, суффиксе и слоге пик приходится на двухфонемные единицы, а в корне - на трехфонемные.

Наиболее общей статистической характеристикой длины единиц языка является их средняя длина, которая вычисляется по формуле:

$$\bar{x} = \frac{\sum x_i n_i}{N},$$

где x_i - количество фонем в исследуемой единице, n_i - частота единиц данной длины, N - количество всех исследуемых единиц.

Полученная нами величина равняется 3,94 фонемы для корня, 2,52 для префикса, 2,4I для слога и 2,35 для суффикса.

Поскольку данная величина зависит от частоты единиц каждой длины, то при вычислении средней длины необходимо установить ее статистические характеристики - среднеквадратичное отклонение /σ/ и меру колебания средней длины /$\sigma_{\bar{x}}$/.

Данные о величине $\bar{x}, \sigma, \sigma_{\bar{x}}, \varepsilon \left(= \frac{2\sigma_{\bar{x}}}{\bar{x}}\right)$ приведены в табл. I.

Таблица I

Средняя длина и ее статистические характеристики

Массив	$\bar{x}$	σ	$\sigma_{\bar{x}}$	ε	$\bar{x} \pm 2\sigma_{\bar{x}}$
Префикс	2,52	0,84	0,0I0	0,008	2,500...2,540
Суффикс	2,35	0,74	0,006	0,005	2,338...2,362
Корень	3,94	I,I4	0,007	0,004	3,926...3,954
Слог	2,4I	0,83	0,002	0,002	2,406...2,4I4

Как и следовало ожидать, самая большая средняя длина характерна для корня. Суффикс оказался самой короткой единицей, а слог занимает промежуточное место между суффиксом и префиксом.

Колебания средней длины исследуемых единиц показывают, что каждая единица выделяется в отдельную группу, что свидетельствует о том, что с точки зрения средней длины префикс, суффикс, корень и слог не представляют собой генеральной совокупности.

Если длина слова зависит от начальной фонемы (Слипченко, 1970), то, очевидно, длина составных его частей (морфем и слогов) тоже будет проявлять такую зависимость.

Для проверки этого предположения каждый исследуемый массив был разбит на совокупности на основании формального критерия начальной фонемы. Полученные совокупности оказались разными по объему: разные фонемы начинают различное количество исследуемых единиц. Это уже свидетельствует о том, что существует определенная зависимость между начальной фонемой и длиной исследуемой единицы.

Зависимоть в распределении морфем и слогов по длине между совокупностями единиц, начинающимися различными фонемами, проявляется также в том, что I) фонемы, которые начинали бы единицы всех длин, отсутствуют во всех массивах; 2) есть фонемы, начинающие единицы только одной длины; 3) гласные не начинают слогов, длиннее четырех, суффиксов - длиннее пяти фонем; 4) для единиц каждой длины имеются свои наиболее характерные фонемы; 5) каждый массив в целом тоже отдает предпочтение каким-то определенным фонемам; 6) единицы преобладающей длины начинаются не всеми, а определенными фонемами (исключение составляют слоги); 7) единицы самой большей длины тоже избирательны в отношении начальной фонемы.

Таким образом, проведенный анализ подтверждает предположение о том, что длина морфемы и слога зависит как от их фонемной структуры, так и от начальной фонемы.

Л И Т Е Р А Т У Р А

Перебийніс В.С. Кількісні та якісні характеристики системи фонем сучасної української літературної мови. - Київ: Наук. думка, 1970. - 272 с.

Слипченко Л.Д. Розподіл довжини слова в англійській мові // Питання структурної лексикології. - Київ: Наук. думка, 1970. - С. 187-197.

Трубецкой Н.С. Основы фонологии. - М.: Изд-во иностр. лит., 1960. - 372 с.

Hocket Ch.F. A Course in Modern Linguistics. - N.Y.: Macmillan, 1958. - 621 p.

ISOMORPHOUS AND DISTINGUISHING FEATURES OF MORPHEMES AND SYLLABLES IN THEIR DISTRIBUTION ACCORDING TO THEIR LENGTH

E.I. Gorot'

S u m m a r y

The article presents the investigation of phonemic structure of four language units - prefix, suffix, root morpheme and syllable. The research made it possible to establish some isomorphous and distinguishing features in the length of compared units.

СИСТЕМЫ АВТОМАТИЗАЦИИ НАУЧНЫХ ИССЛЕДОВАНИЙ В ФИЛОЛОГИИ

А.В. Зубов

Гуманитарные науки отличаются от точных наук следующими особенностями:

1. Неточностью, расплывчатостью своих основных понятий и определений.

2. Преобладанием качественных характеристик их основных объектов.

3. Ограниченностью возможностей проведения активного эксперимента.

4. Большим объемом исходной информации.

Все эти факторы долгое время не позволяли широко использовать ЭВМ для автоматизации научных исследований по гуманитарным наукам и, в частности, в филологии. Однако широкое развитие исследований по проблемам искусственного интеллекта (создание систем машинного перевода, реферирования, аннотирования и поиска, построение экспертных систем) поставило исследователей перед необходимостью создания банков данных и баз знаний, сформированных из слов и словосочетаний естественных языков. Обоснованный отбор таких лингвистических единиц, их всевозможные упорядочения и преобразования требуют большого количества времени значительных коллективов специалистов различного гуманитарного профиля. Во многих случаях такие работы могут быть поручены ЭВМ, способным выполнить их в сотни раз быстрее человека.

Определенный толчок в развитии работ по применению ЭВМ в филологии дала практическая лексикография. Создание различных словарей (толковых, двуязычных, учебных и т.п.) потребовало активного контроля за вновь зарождающейся лексикой. Это можно было сделать лишь с помощью ЭВМ. Используя их и несложный статистический аппарат стало возможным отбирать автоматически лексические минимумы по различным языкам и подъязыкам.

Еще одна причина, способствовавшая созданию и развитию систем автоматизации научных исследований в гуманитарных науках, связана с разработкой новых исследовательских методов (имитационное моделирование, методы статистического и системного анализа и т.п.), которые стало возможным применять и в филологии.

Первый практический шаг в автоматизации исследований в

филологии был сделан в 1952 году, когда американский исследователь Г. Йоссельсон впервые применил ЭВМ для составления частотного словаря русского языка (Автоматизация в лингвистике, 1966). Несколько позже с помощью электронных машин были получены первые результаты по исследованию синтаксических структур английских предложений (США, ЭВМ "SWEAK"). С 1957 года активно использовались ЭВМ во Франции для создания частотных словарей, указателей слов, списков рифм, различных конкордансов, лексических справочников и тезаурусов (Автоматизация в лингвистике, 1966).

В СССР первые опыты по применению ЭВМ в филологии были проведены в 1964–1966 гг., когда с помощью ЭВМ Минск-1, Минск-22 и БЭСМ-3 были получены частотные словари по английским и русским текстам (Лукъяненков, Хотяшов, 1965; Зубов, Хотяшов, 1966; Бородин, 1967).

С возрастанием оперативной памяти и быстродействия ЭВМ они стали особенно широко использоваться в лингвистике и литературоведении.

В лингвистике эксперименты по использованию ЭВМ для получения частотных словарей отдельных текстов переросли в регулярное использование электронной машины при отборе лексических минимумов для обучения различным техническим специальностям[+]. В дальнейшем диапазон применения ЭВМ значительно расширился, и в настоящее время можно выделить следующие направления использования ЭВМ в лингвистике:

1. Построение частотных списков различных лингвистических единиц (словоформ, слов, основ, словосочетаний, предложений).

2. Построение различного типа словоуказателей и конкордансов.

3. Проверка различных лингвистических гипотез о строении и образовании лингвистических единиц.

4. Проверка лингвистических гипотез о строении и образовании совокупностей лингвистических единиц.

5. Исследование законов распределения различных лингвистических единиц и оценка параметров таких распределений.

В современном литературоведении ЭВМ используются:

1. Для изучения стилевых особенностей отдельных авторов.

[+] Перечень таких словарей приведен в работе (Алексеев, 1975).

2. Для изучения ритмических особенностей стихотворных текстов.

3. Для составления словарей рифм.

Наиболее простыми системами в лингвистике являются системы построения частотных списков различных лингвистических единиц (букв, буквосочетаний, слов, словосочетаний, предложений). Такие списки необходимы при отборе наиболее употребительных текстовых единиц в целях создания лексических и грамматических минимусов, словарей и грамматик для систем искусственного интеллекта. Такие списки используются также для теоретического исследования законов распределения лингвистических единиц в тексте и словаре.

Детальный алгоритм выделения различных типов наиболее частых буквосочетаний приведен в работе (Джубанов, 1973). Алгоритм реализован на ЭВМ "Минск-22". Используется предварительно неподготовленный текст.

Известно большое число программ построения частотных списков словоформ. Часть из них реализовано на ЭВМ 2-го поколения. Так в работе (Бектаев и др., 1970) рассмотрены алгоритм и программа для ЭВМ Минск-22 построения частотного словаря казахских словоформ. ЭВМ при этом составляет единый частотный словарь по всему массиву текстов. Аналогичная программа для ЭВМ ЕС-1020 (язык АССЕМБЛЕР) приведена в исследовании (Карпилович, 1973). Программа отлажена на английских текстах. Своеобразный частотный список немецких текстов получен по программе, представленной авторами работы (Вертель, Вертель, 1970). Здесь частотный словарь представлен с учетом длины словоформы. Программа реализована на ЭВМ "Минск-32".

В работах (Бектаев и др., 1970) детально описаны алгоритм и программа построения так называемого обратного частотного словаря, т.е. такого словаря, в котором словоформы упорядочены по алфавиту начиная с конца слова.

Универсальный алгоритм и программа выделения из неподготовленных английских текстов различного типа словосочетаний представлен в исследовании (Долгалева, 1976). Программа написана на языке КОБОЛ и реализована на ЭВМ ЕС-1020.

Ряд систем дистрибутивно-статистического анализа текстов работает с заранее подготовленным текстом. При этом или отдельные классы слов (существительные, прилагательные, глаголы и т.п.) или все они получают до кодирования текста на машинные носители определенные индексы. Например, в иссле-

дования (Архипова, 1980) описана реализованная на ЭВМ ЕС-1022 система статистического анализа грамматических характеристик английских глаголов. С помощью индексов отмечались такие характеристики глаголов как время, лицо, число и т.п. Более широкий объем работ в этом направлении представлен в (Джубанов, 1973). Здесь показана методика выявления с помощью ЭВМ сочетаний отдельных индексов, проставленных заранее в тексте. Такие сочетания могут определять синтаксические, лексические, морфологические и другие признаки слов, обозначенных индексами. С учетом таких же принципов могут быть выделены из массива текстов предложения, содержащие слова определенных классов или сочетания таких слов. Одна из таких систем реализована на ЭВМ Минск-22 (Джубанов, Зубов, 1968). Ряд систем автоматического выделения из текстов слов различных классов, различных сочетаний слов и различных типов предложений представлен в исследовании (Автоматизация анализа ..., 1984). Тексты русских научно-реферативных документов индексируются при этом по указанной выше методике. Программы систем написаны на языках КОБОЛ, ФОРТРАН, языке представления лингвистических знаний (ЯПЛЗ) и реализованы на ЕС ЭВМ.

Еще одна группа систем построения различных списков слов и словосочетаний в качестве основы использует специально построенные тексты или же словари. Так в работе (Бевзенко и др., 1985) представлены результаты работы написанной на языке ПЛ/I программы автоматического составления частотных словарей слов и словосочетаний по поисковым образам документов. Исследование (Ахмеджанов и др., 1984) содержит описание алгоритма построения прямого и обратного частотного словаря словоформ на основе хранящегося в памяти ЭВМ толкового терминологического словаря русского языка. Соответствующие программы реализованы на ЕС-1022. Более сложной является программа построения словаря синонимов эстонского языка на основе хранящихся в памяти ЭВМ толкового и переводных словарей (Viks, Õim, 1985). Программа реализована на ЕС ЭВМ.

В СССР созданы и комплексные системы, позволяющие получать частотные списки различных лингвистических единиц. Одной из таких систем является система, реализованная на ЕС-ЭВМ и ПКВ-М-5000 (Колос, Михайлов, 1977). Эта система состоит из двух групп программ. Группа "СЛОВАРЬ" позволяет строить алфавитные, частотно-алфавитные и ранговые словари для текстов любой длины, вводимых с любого носителя информации. Другая группа - "ЗАПРОС" является информационно-поисковой системой,

выдающей в диалоговом режиме любые комбинации слов, словоформ, словосочетаний и предложений, накопленных в информационном банке системы. Еще более мощная система построена в институте электроники и вычислительной техники АН Латвийской ССР (Якубайтис Т.А. и др., 1980). Она позволяет строить частотные, алфавитно-частотные и обратные словари, искать в массиве текстов слова с определенными грамматическими характеристиками и выдавать списки таких слов, рассчитывать параметры и виды распределения лингвистических единиц, проводить корреляционный и дисперсионный их анализ.

Все большее применение в лингвистике в последние годы находят системы построения различных словоуказателей и конкордансов.

Словоуказатели - это специальные словари слов, которые, как правило, располагаются ЭВМ в алфавитном порядке. При каждом слове таких словарей указывается его общая частота в определенном тексте (или корпусе текстов), число текстов и страниц каждого текста, в которых встретилось слово, и перечень конкретных номеров страниц, на которых встретилось слово, с указанием частоты употребления на данной странице. Иногда указываются и номера строк, на которых встретилось слово. Одна из первых таких систем была реализована на ЭВМ "Минск-32" (Вертель В.А. и др., 1978). При этом исходные данные перфорировались на перфоленту. Получен словоуказатель по поэтическим произведениям С. Есенина. Другая аналогичная система читает исходную информацию с перфокарт (Азарова и др., 1983). С ее помощью получены словоуказатели рассказов М.А. Шолохова, В. Чивилихина, романа М.Ю. Лермонтова "Герой нашего времени", пьесы А.П. Чехова "Вишневый сад". Оригинальная система построения словоуказателя реализована на ЭВМ БЭСМ-6 (Любич и др., 1984). В качестве исходного лингвистического материала здесь использована 6-дорожечная перфолента, которая служила ранее в типографии для фотонаборного печатающего устройства. Программа написана на языке ФОРТРАН-4. С ее помощью получен словоуказатель по книге П.А. Вяземского "Лирика" (М.: Детская литература, 1979. - 183 с.). На ЭВМ ЕС-1022 реализована система построения словоуказателя по текстам казахского писателя М. Ауэзова (Джубанов и др., 1986). В качестве носителя информации использована магнитная лента, подготавливаемая на УПДЛ ЕС-9004. Язык программирования - ПЛ/1.

Более сложной является задача построения конкордансов.

Конкорданс это специфический сложный словарь, который помимо информации, приводимой в описанном выше словаре-словоуказателе, дает для каждого слова контекст, в котором встречается каждое слово в некотором тексте или корпусе текстов. Причем величина контекста колеблется от нескольких слов слева и/или справа от исходного слова текста до нескольких предложений слева и/или справа от такого слова. Первый достаточно большой конкорданс был получен на ЭВМ Минск-32 (Вертель В.А. и др., 1978). Программа читала специально подготовленный текст с перфокарт. Получен конкорданс по произведениям С. Есенина. На перфокартах вводился исходный текст и в программе, построенной для ЭВМ серии ЕС (Азарова и др., 1983). По программе обработаны тексты рассказов М. Шолохова. На языке ПЛ/I написана программа построения конкордансов белорусских авторов (Schastnaya, 1985). Информация готовилась на магнитной ленте (устройство ЕС-9004). Также на магнитной ленте вводилась исходная информация и для построения конкорданса по произведениям казахского писателя М. Ауэзова (Азаев и др., 1982). Соответствующая программа реализована на ЕС ЭВМ.

Единая система построения словоуказателя-конкорданса представлена в исследовании (Марченкова, 1985). Контекст окружения исходного слова в словах определяется здесь границами предложения. Если контекст берется в предложениях, то его глубина может быть от I до IO предложений. Система программ написана на ПЛ/I для ЕС ЭВМ и работает в диалоговом режиме.

Большое число машинных систем создано для проверки различных лингвистических гипотез о строении и образовании различных лингвистических единиц. Часть из них связана с выявлением правил морфологической организации слов. Наиболее многочисленна группа программ, представляющих аналитические модели. Так в работе (Пиотровская, 1977) описывается алгоритм и программа (для ЭВМ 2-го поколения) приведения к канонической (словарной) форме словоупотреблений-существительных и словоупотреблений-прилагательных. Алгоритм основывается на анализе последней буквы существительного и двух последних букв прилагательных. Аналогичная по принципу работы программа для глаголов дана в другом исследовании только что упомянутого автора (Пиотровская, 1973). Более широкая задача решена в исследованиях (Белоногов и др., 1985; Белоногов, Кузнецов, 1983). В первой из них приводятся к канонической форме словоформы и словосочетания из заранее составленных словарей. Во второй входными единицами являются словоупотребле-

ния текста. Программы реализованы на ЕС ЭВМ и опираются на списки двух- и трехбуквенные сочетаний.

Ряд систем направлены на автоматическое выделение из словоформы ее основы. Так в работе (Касаткина, Перепадя, 1976) даны алгоритм и программа на языке "АНАЛИТИК" (ЭВМ "МИР-2") для деления словоформ текста на основу и окончание. Основой алгоритма является список флексий, специфичных для слов различных классов. Предварительно для возможности устранения омонимии слова текста размечаются тремя индексами: существительное, несуществительное, неизменяемая часть речи. К этому же типу систем относится система автоматического деления слова на составные части, представленная в работе (Кобзарова, Лесскис, 1979). Она имеет дело с текстами рефератов по электронике. Предварительно в память ЭВМ вводятся три списка: список основ наиболее употребляющихся в рефератах слов, словарь окончаний и словарь префиксов. На выходе система дает состав слова в виде "префикс + основа + окончание". Программы системы реализованы на ЕС ЭВМ. Более простой системой выделения основ является система, представленная в исследовании (Зубов, 1979). Она опирается на списки одно-, двух-, трех-, четырехбуквенных сочетаний, специфичных для текстов узкой проблемной области "складское хозяйство". Входной единицей является словоформа. На выходе выдается основа с "выброшенными" внутренними гласными. Гласные "выбрасывались" из слов длиной более 8 букв, начиная с 4-5 буквы слова и далее через гласную. Программа написана на языке ПЛ/I и реализована на ЭВМ ЕС-1020. Система, позволяющая восстанавливать исходную словоформу по текстовой, а также объединяющая словоформы в одну парадигму детально описана в исследовании (Морфологический анализ ...).

К описанным системам примыкают и системы автоматического синтеза слов. Так в работе (Коровина, 1975) предлагается алгоритм и программа для синтеза всего ряда словоформ из исходных морфем (используются 118 морфем существительных и 32 морфемы прилагательных). Каждая такая исходная морфема при вводе сопровождается определенным набором лексико-грамматической информации. Программа написана на языке АСТРА и реализована на БЭСМ-6. Своеобразна попытка генерирования киргизской именной словоформы (Мухамедов, Пиотровский, 1986). В ее основу положена А-грамматика Н. Хомского. Эксперимент, проведенный на ЕС ЭВМ позволил получить 300 именных словоформ (из них 288 вполне корректны с точки зрения системы и

нормы киргизского языка). Детальный анализ различных алгоритмов морфологического синтеза, реализованных на ЭВМ, приведен в исследовании (Белоногов, Кузнецов, 1983).

Следующее множество систем автоматизации научных исследований в лингвистике связано с построением программ, проверяющих гипотезы о строении и образовании некоторых совокупностей лингвистических единиц. Одна часть таких систем опирается на качественные признаки лингвистических единиц, другая - на их количественные характеристики. Так в системе, описанной в (Федосимов, Бакулов), путем морфологического анализа слов, происходит их объединение в группы близких по смыслу однокоренных слов. Программы системы написаны на языке ПЛ/I и реализованы на ЕС ЭВМ. Более сложная система представлена в (Анализ метаязыка словаря ..., 1982). Исходной единицей этой системы явилось слово с его определением в толковом словаре. В итоге были получены группы слов, объединенных общими семантическими множителями. В качестве последних выступали повторяющиеся слова, входящие в дефиниции объединенных слов. Реализация осуществлена в рамках автоматизированной системы анализа информации (АСАИ), разработанной в институте социологических исследований АН СССР (Молчанов, Афанасьев, 1978). К числу систем, использующих количественные характеристики слов можно отнести систему (Мамедова, Скороходько, 1981), опирающуюся на хранящийся в памяти ЭВМ терминологический словарь, в котором каждый термин сопровождается его дефиницией (до 20 слов). В составе дефиниций специальными метками выделяются слова, рассматриваемые в качестве непосредственных семантических составляющих определяемого термина. Система позволяет проводить различные группировки исходных терминов (по наличию семантических составляющих, по признаку связности терминов и др.), а также вычислять статистические параметры отдельных слов и всего словаря. Последние подсчеты осуществляются по специальным формулам. Система реализована на ЕС ЭВМ и АСВТ-М (язык ПЛ/I). Более простая система представлена в (Перцев, Румшинский). Она позволяет из неподготовленного текста отобрать наиболее информативные основы. Для такого отбора используется разница распределений слов в двух частотных словарях, полученных на далеких друг от друга подъязыках. Программы системы реализованы на ЭВМ "Ирис-50" на материале рефератов по теме "Электрические машины и аппараты". В работе (Зубов, Чапля, Чапля) представлена система, позволя-

ющая разделить массив исходных терминов на группы ключевых слов по их принадлежности к определенным темам определенной предметной области. Основанием для такого разделения послужила формула условной вероятности Бейеса и коэффициент вариации слова, вычисляемые для каждого термина на основе частот употребления последнего во всех темах предметной области. Система реализована на ЭВМ "Минск-22". Сходная процедура использована в работе (Gurova, 1980) для выделения из частотного словаря некоторого базового словаря, покрывающего до 88% слов любого текста определенной специальности (тексты по металлургии). Для такого выделения используются коэффициенты вариации и дисперсии, а также коэффициенты употребительности и распространенности слова. Программы реализованы на БЭСМ-4. Впервые реализован с помощью ЭВМ в лингвистике кластерный анализ (Солнцев, 1986). В этой работе он применен для разделения массива прилагательных подъязыка английской вычислительной техники на семантические группы и подгруппы. Программа написана на языке ПЛ/I и реализована на ЭВМ ЕС-1035.

Последняя группа систем автоматизации научных исследований в лингвистике связана с исследованием законов распределения различных лингвистических единиц в текстах и словарях и оценкой параметров таких распределений. Одной из наиболее значительных работ этого типа было исследование (Бектаев, Лукьяненков, 1971). В нем проверялись эмпирические распределения 300 словоформ и 300 трехсловных сочетаний взятых из английских текстов подъязыка судовых механизмов. Эти распределения проверялись относительно следующих теоретических распределений: нормального закона, логарифмически-нормального и распределения Пуассона. Все вычисления и оценки проведены на ЭВМ Минск-22. Несколько ранее было проведено исследование по анализу законов распределения трехсловных сочетаний в немецком подъязыке публицистики и английских текстах по электронике (Бектаев и др., 1969). Оценка близости эмпирических и теоретических распределений проводилась с помощью критерия x^2-квадрат. Проверялось соответствие эмпирических законов закону Пуассона и биноминальному закону. Система программ реализована на ЭВМ "Минск-2". К такому же типу систем относится и система, представленная в (Манасян, 1981). В ней с помощью ЭВМ ЕС-1020 проводился анализ распределений 1025 терминов в английском подъязыке квантовой механики и электроники на соответствие законов их

распределения законам Пуассона, нормальному и логнормальному. Одна из последних систем такого типа реализована на ЕС-1035 (язык ПЛ/I). В ней осуществлялся поиск законов распределения в тексте украинских слов, взятых из текстов различных писателей (Марченкова, 1986). Проверялось с помощью критерия x^2-квадрат соответствие эмпирических распределений этих слов закону Пуассона, нормальному закону и отрицательному биноминальному закону. К этому же типу систем относится и система (Тулдава, 1986), исследуемая законы распределения слов по их длине в текстах и в словаре (длина измерялась и в буквах и в слогах). Изучалось соответствие логнормальному закону, закону Чебанова-Фукса и логарифмическому закону на эстонском материале. Программы реализованы на ЕС ЭВМ. Более широкая система анализа типов эмпирических распределений и оценки их параметров представлена в работе (Милюна, Якубайтис, 1980). На материале латышских текстов различного типа (художественных, научно-технических, газетно-журнальных и деловых) исследовались эмпирические распределения слов этих текстов на соответствие их следующим статистическим законам: нормальному, логнормальному, гамма, бета, Вейбулла. При этом оценивается и адекватность подобранных моделей. Система реализована на ЕС ЭВМ. Вплотную примыкают к описанным системам комплекс программ, представленный в (Бахмутова, 1986). Его цель - автоматическая проверка признаков качественной однородности 6-ти английских подъязыков. Под качественной однородностью автор понимает идентичность статистического и графического "поведения" выбранных подъязыков в рамках модели Ципфа. Вся система реализована на языке ФОРТРАН с использованием графопостроителя.

Известна попытка применения ЭВМ для проведения корреляционного и факторного анализа (Тулдава, 1976). Корреляционный анализ проводился с целью выявления взаимозависимости 12-ти стилеметрических параметров (номинальность, частота употребления прилагательных, личных местоимений, спрягаемых глаголов и т.п.) в текстах семи эстонских писателей. Полученные при этом парные коэффициенты корреляции использованы далее для проведения факторного анализа. При этом ЭВМ "Минск-32" выделила четыре главных фактора.

В работах (Зубов, Хотяшов, 1966; Беляева, 1974) приводятся алгоритмы и программы вычисления параметров K, β, γ вероятностного закона, известного в лингвистике под именем закона Эсту-Ципфа-Мандельброта. В обоих случаях для этого

использовался метод наименьших квадратов. Программы первой системы реализованы на "Минск-2", а второй - на ЭВМ "Одра-1204" (язык АЛГОЛ-60). В первой из упомянутых двух работ дан также алгоритм для вычисления методом наименьших квадратов параметров и интегрального закона распределения вероятностей.

Менее многочисленны системы автоматизации научных исследований в литературоведении.

Среди систем, описывающих статистические особенности отдельных авторов можно выделить реализованную на ЭВМ БЭСМ-3М систему (Мальцева, 1969). Она исследовала русские эпистолярные тексты на основе 22-х параметров. В основе алгоритма системы лежат различные частотные словари, полученные по 160 частным письмам и вычисляемые на их основе различные статистические критерии. Система описания индивидуальных авторских особенностей С. Есенина представлена в исследовании (Зубов, 1981). В ее основе лежит словарь поэтических произведений С. Есенина (Гайдукова, Зубов, 1975), в котором каждой словоформе дан ряд индексов (6 знаков), характеризующих лексико-грамматические и структурные особенности словоформы. Система позволяет выделить типичные для поэта классы слов, грамматические категории, грамматические и ритмические структуры строк. Программы системы написаны на языке ПЛ/1 и реализованы на ЭВМ ЕС-1020.

Удачный опыт применения ЭВМ для классификации текстов по индивидуально-авторским особенностям приведен в работе (Тулдава, 1981). Здесь впервые в СССР в филологических целях использован кластерный анализ. Известна попытка автоматического определения авторов средневековых текстов (Бородкин, 1986).

Ряд систем связан с анализом ритмической структуры поэтических текстов. В работе (Баевский, Осипов, 1974) изучались стихи альтернирующего ритма (68 стихотворений, написанных пятистопным хореем на русском языке). Предварительно отмечались индексами: ударность слога, степень его выделенности в языке, положение на сильном или слабом метрическом месте стиха. Подсчет средних арифметических выделенности слогов позволял описать ритмическое своеобразие 16 поэтов. Система программ реализована на ЭВМ "Минск-32". В работах (Rüütel, 1981; Рюйтел, 1984) приведена реализованная на ЭВМ ЕС-1010 система автоматического статистического анализа распределения связей отдельных элементов ме-

лодии, т.е. мелодического контекста. Нормативные модели сравниваются с исследуемыми эстонскими руническими напевами. Такое сравнение позволило провести типологическую классификацию этих напевов. Близость напевов определялась формулой Чурпова. Процедура, имитирующая процесс регрессивной акцентной диссимиляции в XII веке была смоделирования на БЭСМ-3М в работе (Красноперова, 1982). Построенная модель вполне достоверно предсказала вероятности появления ритмических форм в последующие периоды, имея данные о предшествующих. Система, рассматривающая стихотворение как временной ряд, описана в (Васюточкин, 1980). Для исследования такого ряда использован спектральный анализ. Использованы стихи Блока, Есенина, М. Кузьмина и других поэтов. Система первоначально реализована на ЭВМ "Мир", а позже - на М-222. Программа предусматривала получение не только основных спектральных характеристик, но и представление спектрограмм в любом - натуральном, логарифмическом или полулогарифмическом масштабе.

Ряд систем подобного типа направлены на выявление статистической связи звука и смысла в стихах. Так, в исследовании (Журавлев, 1969) рассмотрена система, реализованная на ЭВМ "Минск-22", оценивающая эмоциональный фон стихотворного текста. Используются 34 признака фонотипа с помощью методов корреляционного анализа. Анализу подвергались стихи Маяковского, Блока, Есенина и других поэтов. Особенно ярко указанное направление проявилось в исследованиях (Красноперова, 1974; Красноперова, Боголюбова, 1979). Здесь описана реализованная с помощью ЭВМ БЭСМ-3М система, выявляющая зависимость между ритмическими структурами текстов и их смыслями. Исследовались как поэтические, так и прозаические художественные тексты. Основной "разрешающий" критерий - частота употребления ритмических структур.

Важная и трудоемкая для литературоведов задача составления словаря рифм с помощью ЭВМ пока не находит практической реализации в нашей стране. Детальный алгоритм построения словаря русских рифм с помощью ЭВМ приведен в работе (Левин, 1980).

ЛИТЕРАТУРА

Автоматизация анализа научного текста. - Киев: Наукова думка, 1984. - 257 с.

Автоматизация в лингвистике. - М.-Л.: Наука, 1966. - 157 с.

Азаев Х.Г., Бектаев К.Б., Валиев Х.Ш. и др. Использование ЭВМ в авторской лексикографии // Переработка текста методами инженерной лингвистики. Тезисы докладов. Всесоюзная конференция. - Минск, 1982. - С. 137.

Азарова И.В., Горохова С.И., Григорьев Г.Г., Кузнецова Е.Л. Разработка автоматических словоуказателей и конкордансов для художественных текстов // Структурная и прикладная лингвистика. Вып. 2. - Л.: ЛГУ, 1983. - С. 187-190.

Алексеев П.М. Статистическая лексикография: типология, составление и применение частотных словарей. Учебное пособие. - Л.: ЛГПИ, 1975. - 120 с.

Анализ метаязыка словаря с помощью ЭВМ. - М.: Наука, 1982. - 95 с.

Архипова Т.И. Опыт статистической обработки текстов в целях отбора грамматического минимума // Материалы семинара Статистическая оптимизация преподавания языков и инженерная лингвистика. - Чимкент, 1980. - С. 151-152.

Ахмеджанов М.М., Гельфман Г.С., Королев Э.И., Мальцев Е.П. Автоматизированная словарно-терминологическая службы по системам связи // НТИ, сер. 2, 1984, № 12. - С. 23-33.

Баевский В.С., Осипова Л.Я. Исследование стихотворного ритма с использованием ЭВМ // Структурная и математическая лингвистика. Вып. 2. - Киев, 1974. - С. 11-19.

Бевзенко Е.А., Гальперина Т.А., Крученицкая Е.А., Кузнецов Б.А., Хорошилов А.А. Автоматическое составление частотных словарей ключевых слов и словосочетаний по поисковым образам документов // НТИ, сер. 2, 1985, № 8. - С. 28-31.

Бектаев К.Б., Зубов А.В., Королевич Е.Ф. и др. К исследованию законов распределения лингвистических единиц // Статистика текста. Том 1. Лингвистические исследования. - Минск, 1969. - С. 131-162.

Бектаев К.Б., Джубанов А.К., Зубов А.В. Автоматическое построение частотных словарей (прямого и обратного) // Вестник Академии наук Казахской ССР. № 3 (299), март 1970. - С. 48-53.

Бектаев К.Б., Лукъяненков К.Ф. О законах распределения единиц письменной речи // Статистика речи и автоматический анализ текста. - Л.: Наука, 1971. - С. 47-112.

Белоногов Г.Г., Загика Е.А., Новоселов А.П. и др. Автоматическая нормализация слов и словосочетаний // НТИ, сер. 2, 1985, № 1. - С. 37-40.

Белоногов Г.Г., Кузнецов Б.А. Языковые средства автоматизированных информационных систем. – М.: Наука, 1983. – С. 146–183.

Бестужев А.К., Городецкий Б.Ю., Зайцева О.В. и др. Семантико-квантитативные исследования подъязыка (опыт создания автоматизированной системы) // Квантитативная лингвистика и автоматический анализ текстов. Уч. зап. Тартуск. гос. унив. Вып. 745. – Тарту, 1986. – С. 37–49.

Автоматизация лексикографических работ // Научные труды НИИ ПМК. – Горький, 1967. – С. 33–41.

Бородкин Л.И. ЭВМ ищет авторов средневековых текстов // Число и мысль. Выпуск 9. – М.: Знание, 1986. – С. 113–141.

Васюточкин Г.С. О некоторых аспектах применения математики и ЭВМ в стиховедении // НТР и развитие художественного творчества. – Л.: Наука, 1980. – С. 202–216.

Вертель В.А., Вертель Е.В., Рогожникова Р.П. К вопросу об автоматизации лексикографических работ (некоторые результаты применения ЭВМ) // Вопросы языкознания, 1978, № 2. – С. 104–110.

Гайдуков З.С., Зубов А.В. Частотный словарь поэтических произведений С. Есенина // Вопросы общей и прикладной лингвистики. – Минск, 1975. – С. 165–186.

Джубанов А.Х. Статистическое исследование казахского текста с применением ЭВМ (на материале романа Н. Ауэзова "Абай жолы"). Автореф. дис. ... канд. филол. наук. – Алма-Ата, 1973. – 37 с.

Джубанов А.Х., Алдабергенова А., Белботаев А., Зекенова А. Опыт построения на ЭВМ словаря-словоуказателя автора // Проблемы автоматического и экспериментально-фонетического анализа текстов. – Минск: Изд-во Минского ГПИИЯ, 1986. – С. 146–150.

Джубанов А.Х., Зубов А.В. Автоматизация некоторых лингвистических процессов // Вестник Академии наук Казахской ССР 1968, № 9 (280), сентябрь. – С. 31–36.

Долгалева Т.В. Программа выделения дистрибутивных словосочетаний для ЭВМ ЕС-1020 // Автоматический анализ текстов. – Минск, 1976. – С. 184–189.

Журавлев А.П. Автоматический анализ эмоционального фона стихотворного текста // Проблемы прикладной лингвистики. Тезисы межвузовской конференции 16–19 декабря 1969 г. Часть I. – М.: 1969. – С. 119–121.

Зубов А.В. Автоматическая компрессия терминологической части тезауруса // Романское и германское языкознание. Вып I. Вопросы экспериментальной фонетики и прикладной лингвистики. - Минск, I979. - С. I92-I96.

Зубов А.В. Автоматический статистический анализ поэтического текста // Актуальные проблемы квантитативной лингвистики и автоматического анализа текстов. Труды по лингвостатистике: Уч. зап. Тартуск. гос. унив. Вып. 59I. - Тарту, I98I. - С. 35-45.

Зубов А.В., Хотяшов Э.Н. Статистический анализ текста с помощью электронно-вычислительной машины // Энтропия языка и статистика речи. - Минск: Изд-во Минского ГПИИЯ, I966. - С. II8-I66.

Зубов А.В., Чапля А.И., Чапля С.Г. Автоматическое выделение ключевых слов // Структурная и прикладная лингвистика. Вып. I. - Л.: ЛГУ, I978. - С. I98-204.

Карпилович Т.П. Программа построения частотно-алфавитного словаря на машине с байтовой структурой // Лингвистика и автоматический анализ текстов. - Минск, I973. - С. 450-459.

Касаткина И.В., Перепадя И.М. Плгоритм выделения флексий русского языка // Структурная и математическая лингвистика. Вып. 4. - Киев: Вища школа, I976. - С. 89-92.

Кобзарева Т.Ю., Лесокис Г.А. Система автоматического морфологического членения текста // НТИ, сер. 2, I979, № 2. - С. 23-27.

Колос Г.И., Михайлов В.А. Комплекс программ ЕС-ЭВМ и ПВК М-5000 для обработки текстов // Автоматическая переработка текста методами прикладной лингвистики. Материалы П Всесоюзной конференции 6-7 октября I977 г. Кишинев, I977. - С. I04-I06.

Красноперова М.А. Модель восприятия и порождения ритма // Вопросы кибернетики. - М.: I982. - С. I24-I40.

Красноперова М.А. Об использовании ЭВМ для статистического анализа стихотворных текстов // Методы вычислений. Вып. 9. - Л., I974.

Красноперова М.А., Боголюбова Н.А. Испытание на связь между ритмическим эффектом и смыслом // Лингвистические проблемы функционального моделирования речевой деятельности. Вып. IУ. - Л., I979. - С. I75-I93.

Левин М.Ю. Применение ЭВМ для составления словаря рифм // НТР и развитие художественного творчества. - Л.: Наука, I980. - С. I96-202.

Лукъяненков К.Ф., Хотяшов Э.Н. Составление частотно-алфавитного словаря с помощью электронно-вычислительной машины // Вопросы романо-германской филологии и статистика речи. Часть I. - Минск: Издво Минского ГПИИЯ, 1965. - С. 77-100.

Любич Д.В., Рогожникова Р.П., Чернышева Л.В. Получение словоуказателя на ЭВМ // Теория и практика современной лексикографии. - Л.: Наука, 1984. - С. 38-76.

Мамедова М.Г., Скороходько Э.Ф. Автоматизированная система анализа терминологической лексики // НТИ, сер. 2, 1981, № 1. - С. 14-18.

Мальцева Г.Ф. Некоторые количественные приемы описания индивидуального авторского стиля // Статистика текста. Том I. Лингвостатистические исследования. - Минск, 1969. - С. 206-247.

Манасян Н. О распределениях терминов в английском научно-техническом тексте (подъязык квантовых генераторов) // Актуальные проблемы квантитативной лингвистики и автоматического анализа текстов: Уч. зап. Тартуск. гос. унив. Вып. 591. - Тарту, 1981. - С. 60-73.

Марченкова Е.Л. Об одном алгоритме построения словоуказателя-конкорданса // Системные характеристики устной и письменной речи. - Минск, 1985. - С. 193-199.

Марченкова Е.Л. Программа вычисления некоторых параметров законов распределения лингвистических единиц // Проблемы автоматического и экспериментально-фонетического анализа текстов. - Минск, 1986. - С. 186-191.

Милюна М., Якубайтис Т. Автоматизированная система подбора вида распределения лингвистических единиц // Лингвостатистика и квантитативные закономерности текста. Труды по лингвостатистике: Уч. зап. Тартуск. гос. унив. Вып. 549. - Тарту, 1980. - С. 86-105.

Морфологический анализ научного текста на ЭВМ. - Киев: Наукова думка, 1989. - 261 с.

Мухамедов С.А., Пиотровский Р.Г. Инженерная лингвистика и опыт системно-статистического исследования узбекских текстов. - Ташкент, 1986. - 161 с.

Перцев Л.Г., Румшинский Б.Л. Автоматическое построение словника для ИПЯ дескриторного типа // Автоматическая переработка текста методами прикладной лингвистики. Материалы П Всесоюзной конференции 6-7 октября 1977 года. - Кишинев, 1977. - С. 175-176.

Пиотровская А.А. Автоматическое приведение именных словоупотреблений к канонической форме // НТИ, сер. 2 1977, № 1. - С. 32-36.

Пиотровская А.А. Машинная морфология русского глагола // Статистика речи и автоматический анализ текста. - Л.; 1973. - С. 260-277.

Рюйтел И. Опыт определения мелодических типов и их взаимосвязей с иными признаками (на материале эстонских рунических напевов) // Квантитативная лингвистика и автоматический анализ текстов. Труды по лингвистике: Уч. зап. Тартуск. унив. Вып. 689. - Тарту, 1984. - С. 133-146.

Солнцев В.П. Синтагматический подход к выделению групп прилагательных // Проблемы автоматического и экспериментально-фонетического анализа текстов. - Минск, 1986. - С. 201-206.

Тулдава Ю.А. Длина слова и распределение слов по длине в тексте и в словаре // Исследования по общему и сопоставительному языкознанию: Уч. зап. Тартуск. гос. унив. Вып. 736. - Тарту, 1986. - С. 150-166.

Тулдава Ю.А. Опыт квантитативного анализа художественного стиля // Уч. зап. Тартуск. гос. унив. Вып. 396. - Тарту, 1976. - С. 122-141.

Тулдава Ю. Опыт классификации текстов с помощью кластерного анализа // Уч. зап. Тартуск. гос. унив. Вып. 591. - Тарту, 1981. - С. 136-157.

Федосимов И., Бакулов А.Д. Морфемный анализ словарных форм // НТИ, сер. 2, 1979, № 4. - С. 27-30.

Якубайтис Т.А., Удалов В.И., Склярович А.Н. Лингвостатистические исследования с использованием экспериментальной вычислительной сети АН Латвийской ССР // Материалы семинара "Статистическая оптимизация преподавания языков и инженерная лингвистика". - Чимкент, 1980. - С. 30-31.

Gurova, N.V. Computational parameters of basic vocabulary selection from a word count // Symposium:Computational linguistics and related topics (Tallinn, November 24-26, 1980). Summaries. - Tallinn, 1980, p. 52-53.

Rüütel, I. Typology of Estonian runo-tunes: experiment and some results. Preprint KKI-18. - Tallinn, 1981.

Schastnaya, N.M. Principles of automated compilation of a concordance to a certain language // Symposium of automatic compilation of dictionaries (Tallinn, November

25-27, 1985). Summaries. - Tallinn, 1985, p. 59-60.

Viks, Ü. An attempt to outline the structure of a data base of dictionaries // Symposium on automatic compilation of dictionaries (Tallinn, November 25-27, 1985). Summaries. - Tallinn, 1985, p. 82-83.

Viks, Ü., Õim, A. Computer-aided compilation of the dictionary of estonian synonyms // Symposium on automatic compilation of dictionaries (Tallinn, November 25-27, 1985). Summaries. - Tallinn, 1985, p. 88-90.

A SYSTEM OF AUTOMATING SCIENTIFIC RESEARCH IN PHILOLOGY

Alexander Zubov

S u m m a r y

The article contains chronological data on the computer algorithms and programs used for linguistic and belletristic analysis of texts in this country. The modern computer capabilities are demonstrated for compiling various kinds of dictionaries and for testing various statistical hypotheses on the distribution of individual linguistic units and their aggregates.

ИЗБИРАТЕЛЬНОСТЬ СОЧЕТАНИЯ СМЫСЛОВ И ВОЗМОЖНЫЕ СПОСОБЫ ЕЕ СТАТИСТИЧЕСКОГО ВЫРАЖЕНИЯ

В.Ю. Иванюк, В.В. Левицкий

Интенсивное исследование лексической сочетаемости привело к открытию и использованию важного понятия, обозначаемого чаще всего терминами "селективность" или "избирательность". Способностью сочетаться или не сочетаться друг с другом обладают, однако не только слова, но и обозначаемые ими сегменты содержания, т.е. "смыслы", на что в свое время указывал еще Л.В. Щерба. Сочетаться друг с другом могут как "лексические", так и "грамматические" смыслы. При этом возможны три основные комбинации: лексический плюс лексический смысл (например, "цвет" + "зависть"); грамматический + грамматический (например, "время" + "лицо"); лексический плюс грамматический (например, "Тартуский университет" + "множественное число"). Способность данного смысла сочетаться или, наоборот, не сочетаться с определенным набором других смыслов будем называть избирательностью сочетаемости смыслов. В отличие от лексической сочетаемости, которой обозначают соединение слов, сочетание смыслов целесообразно называть семантической сочетаемостью.

Возникает вопрос: поддается ли понятие "семантическая сочетаемость" статистическому анализу и, если да, то каким образом может быть более или менее точно выражена степень семантической совместимости? Попытаемся найти возможные пути решения этой задачи с помощью экспериментального исследования взаимодействия лексических и грамматических значений (смыслов) в сфере употребления грамматических времен немецкого языка. Сформулируем в самом общем виде вопрос, на который предстоит ответить, следующим образом: носит ли употребление глаголов с определенной лексической семантикой в том или ином грамматическом времени произвольный характер или подчиняется каким-то закономерностям?

Прежде всего необходимо, очевидно, выделить сами единицы анализа, т.е. то, что обозначено выше термином "смысл". Выделение инвентаря лексических или грамматических значений возможно либо путем интуитивно-логического разбиения семантического континуума на некоторое число сегментов, либо путем классификации и группировки по заданным признакам тех или иных формальных единиц. Например, при разбиении семантики слова

на "значения" можно определенным образом сгруппировать его контекстуальные наборы (партнеры) и по числу и характеру полученных группировок судить о числе и содержательном наполнении лексико-семантических вариантов слова. При разбиении более крупных смысловых участков можно сгруппировать в лексические микросистемы всю анализируемую лексику, т.е. в конечном счете разбить некоторое исследуемое множество на подмножества. Разумеется, и при использовании "чисто" логического анализа исследователь в той или иной степени опирается не только на содержание, но и на форму или употребление исследуемых семантических единиц.

Воспользуемся для выполнения запланированного эксперимента обоими способами получения инвентаря смыслов. С учетом существующих в германистике классификаций (имеется в виду прежде всего работа Helbig, Buscha, 1986) весь семантический спектр грамматических времен в немецком языке был разбит нами на следующие смысловые сегменты: 1) актуальный презенс; 2) генеральный презенс; 3) футуральный презенс; 4) исторический презенс; 5) претерит; 6) перфект для обозначения прошедшего; 7) результативный перфект; 8) футуральный перфект; 9) плюсквамперфект; 10) футур I. Кроме того, были выделены еще некоторые темпоральные смыслы, которые,как показало дальнейшее исследование, обладают крайне низкой частотой встречаемости (футур II, модальный футур I и др.).

Лексические смыслы были получены путем классификации глаголов. Все анализируемые глаголы были разбиты на 4 крупных класса: глаголы состояния, процесса, действия и модальные глаголы (ср. Helbig, 1983, S. 67-105). Четыре полученных таким образом семантических сегмента были обозначены как "состояние", "процесс", "действие", "модальность". Эти сегменты и составляют инвентарь лексических смыслов. Более мелкие лексические смыслы, полученные путем дальнейшей классификации глаголов (разбиения классов на подклассы), в данной статье не рассматриваются.

Материалом исследования послужили 9043 микроконтекста, содержащих глаголную синтагму и полученных путем сплошной выборки из текстов 4 функциональных стилей (художественная проза, драматургия, публицистика, научный текст). В итоге зафиксировано употребление глаголов: в презенсе - 4200, в претерите - 3037, перфекте - 1105, плюсквамперфекте - 417, футуре I - 282, футуре II - 2.

Таблица I

Частота совместной встречаемости лексических и грамматических смыслов

грамматическая семантика	лексическая семантика					всего
	состояние	процесс	действие	модальность	прочие	
актуальный презенс	+923	70	557	+404	27	I98I
генеральный презенс	+8I8	+I75	3I2	+I59	29	I493
исторический презенс	48	I7	+I4I	6	2	2I4
футуральный презенс	I0I	+76	+280	40	I5	5I2
претерит	I047	285	+I50I	I70	34	3037
перфект прошедшего	I22	63	+542	I7	3I	775
результативный перфект	22	+84	+I90	0	I3	309
плюсквамперфект	79	+79	+24I	5	I3	4I7
футур I	67	23	+I50	8	I8	266
прочие	II	7	I5	0	6	39
в с е г о	3238	879	3929	809	I88	9043

Знаком "+" в таблице обозначены частоты, статистически существенно превышающие теоретически ожидаемые величины.

Распределение частот сочетаемости лексических и грамматических смыслов отражено в таблице I (некоторые низкочастотные разряды грамматических смыслов в целях экономии места объединены; в лексический класс "прочие" вошли значения глаголов в составе фразеологизмов). Как видно из табл. I, частоты совместной встречаемости лексических и грамматических смыслов распределяются неравномерно. Такая неравномерность зависит, очевидно, от трех основных факторов: а) от неодинаковой частоты употребления различных грамматических времен; б) от неодинаковой частоты употребления различных семантических классов и подклассов слов (в данном случае - глаголов с различной семантикой); в) от сочетаемостных свойств представленных в таблице смыслов, т.е. именно от того фактора, который обозначен выше как способность одних смыслов сочетаться или не сочетаться с другими и назван избирательностью сочетаемости. Следовательно, для того, чтобы измерить степень избирательности сочетаемости смыслов, необходимо найти такую статистическую процедуру, которая бы нейтрализовала действие двух первых факторов. Такой процедурой могло бы быть статистическое изучение соответствия

эмпирических частот теоретическим с помощью критерия χ^2. Теоретические частоты рассчитываются в этом случае пропорционально маргинальным суммарным частотам по столбцам и строкам таблицы, т.е., переводя в плоскость нашего исследования, пропорционально частотам, обусловленным действием двух первых указанных факторов (частоты употребления лексических и грамматических смыслов). Если эмпирические частоты будут соответствовать теоретическим и максимально приближаться к ним, то это будет означать, что наблюдаемое распределение частот подчиняется воздействию только первых двух факторов и зависит главным образом или исключительно от частоты употребления смыслов, а не от их совместимости / несовместимости. Если же, наоборот, в одних случаях наблюдаемые частоты будут существенно превышать ожидаемые, а в других будут значительно ниже таковых, то это будет свидетельствовать о некотором "сверхпритяжении" или некоторой аномальной несовместимости смыслов, обусловленных действием селективных процессов, протекающих в языке, но скрытых от внешнего наблюдения. Статистическую процедуру необходимо построить таким образом, чтобы получить возможность измерить сочетательные способности того или иного смысла с другим смыслом (попарно) или с группой (набором) смыслов.

Для осуществления такого построения процедуры составим на основе многопольной таблицы I четырехпольные (альтернативные) таблицы. В качестве примера приводится таблица 2.

Таблица 2

Распределение частот сочетаемости актуального презенса

лексическая семантика	грамматическая семантика		всего
	актуальный презенс	другие смыслы	
состояние	923	2315	3238
другие смыслы	1058	4747	5805
в с е г о	1981	7062	9043

Вычисление χ^2 проводится по формуле:

$$\chi^2 = \frac{(ad - bc)^2 \cdot N}{(a + b)(c + d)(a + c)(b + d)}, \qquad (I)$$

где a,b,c,d числа в полях четырехпольной таблицы, а N сумма всех наблюдений (в нашем примере N = 9043).

Соответствующие расчеты показывают, что для табл. 2 χ^2 = I28. Однако сама по себе величина χ^2 свидетельствует в данном случае лишь о том, что эмпирические частоты существенно отличаются от теоретических, но она не может служить мерой соответствия одних частот другим, т.е. - в нашем случае - мерой избирательности сочетаемости. В качестве такой меры может быть использован коэффициент взаимной сопряженности А.А. Чупрова, в формулу которого входит величина χ^2 (см. Рокицкий, I964, с. 257; Урбах, I964, с. 359):

$$K = \sqrt{\frac{\chi^2}{N\sqrt{(r-1)(c-1)}}}, \qquad (2)$$

где r и c число рядов и столбцов таблицы. Для табл. 2 величина K составляет 0,II9. Коэффициент сопряженности может быть найден и без вычисления χ^2 по формуле:

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \qquad (3)$$

Однако мы отдаем предпочтение двухступенчатой процедуре: сначала вычисляется величина χ^2, затем - коэффициент сопряженности. В этом случае исследователь получает возможность вначале определить, существенно или не существенно отклоняются эмпирические частоты от теоретических, а затем - в зависимости от полученных результатов - переходить к нахождению коэффициента сопряженности. Естественно, если χ^2 не обладает необходимой значимостью (при df = I и P = 0,05 сумма χ^2 не должна быть меньше 3,84), вычисление величины коэффициента сопряженности теряет смысл. Формула (3) имеет, правда, некоторые преимущества, так как позволяет определить знак коэффициента ("плюс" или "минус"). Знак "плюс" свидетельствует о повышенной совместимости, а знак "минус" - об аномальной несовместимости семантических единиц. Однако и при вычислении величины χ^2 по формуле (I) о направлении сопряженности можно легко судить по разнице (ad - bc). Если эта разница положительна, значит и коэффициент сопряженности, вычисленный по формуле (3), будет положительным.

Об избирательности сочетаемости грамматической смысловой единицы с набором лексических смыслов можно судить по среднему коэффициенту сопряженности (принимаются в расчет абсолютные величины коэффициентов). Так, для актуального презенса получены коэффициенты: +0,II9 (в сочетании со смыс-

лом "состояние"); -0,044 ("процесс"); -0,164 ("действие") ; +0,212 ("модальность"). Средний коэффициент равен 0,134 (по абсолютной величине). Соответственно для исторического презенса средний коэффициент К = 0,047, для плюсквамперфекта К = 0,067. Следовательно, можно полагать, что актуальный презенс обладает большей избирательностью, чем плюсквамперфект или исторический презенс. Разумеется, необходимые коэффициенты можно получить и другим путем: сравнивая частоты одной семантической единицы (например, актуального презенса) с частотами всех других единиц с помощью многопольных таблиц.

Мера избирательности сочетаемости каждого из классов глаголов также легко находится путем суммирования соответствующих коэффициентов. По степени избирательности своей сочетаемости исследованные 4 класса глаголов распределяются следующим образом: глаголы действия (К = 0,102); глаголы состояния (К = 0,089); модальные глаголы (К = 0,074); глаголы процесса (К = 0,055). При этом глаголы действия, например, аномально несовместимы с актуальным и генеральным презенсом, и, наоборот, обладают повышенной совместимостью с перфектом прошедшего, претеритом, историческим презенсом. Следует еще раз подчеркнуть, что величина коэффициента К не зависит от частоты совместной встречаемости смыслов. Так, частота сочетания смыслов "процесс" и "претерит" (285) приближается к некотрой "норме", пределы которой обусловлены только частотой употребления каждой из двух составляющих - глаголов процесса и прошедшего времени претерит. К такой же норме приближается частота сочетания смыслов "претерит" и "состояние" (1074), "исторический презенс" и "процесс", "модальность" и "футуральный презенс". Во всех этих случаях эмпирические частоты почти не отличаются от теоретических.

Таким образом, отвечая на поставленный выше вопрос (с. 55), следует полагать, что употребление различных семантических классов глаголов в том или ином грамматическом времени подчиняется определенным закономерностям и носит специфический характер.

Как показал Ю.А. Тулдава (1988, с. 157-158), при сравнении результатов различных экспериментов целесообразнее использовать так называемые нормированные коэффициенты. Возможно, применение этих коэффициентов внесет некоторые коррективы в полученные нами данные. Однако цель нашего исследования, как явствует из его заглавия, заключается не столь-

ко в получении конкретных данных (хотя, разумеется, они также представляют интерес), сколько в поисках объективных способов измерения сочетательных потенций семантических единиц. Предложенная процедура может быть использована и при изучении лексической сочетаемости (см. Левицкий, Быстрова, Гинка, 1987).

Л И Т Е Р А Т У Р А

Левицкий В.В., Быстрова Л.В., Гинка Б.И. Изучение лексической сочетаемости с помощью статистических методов // Функционирование языковых единиц в речи и тексте: Межвузовский сборник научных трудов. - Воронеж: Изд-во Воронеж. ун-та, 1987. - С. 48-59.

Рокицкий П.Ф. Биологическая статистика. - Минск: Высшая школа, 1964. - 412 с.

Тулдава Ю.А. Об измерении связи качественных признаков в лингвистике (I): Сопряженность альтернативных признаков // Квантитативная лингвистика и автоматический анализ текстов. Уч. зап. Тартуского ун-та, 827. - Тарту, 1988. - С. 146-162.

Урбах В.Ю. Биометрические методы. - М.: Наука, 1964. - 412 с

Helbig G. Studien zur deutschen Syntax. - Leipzig: VEB Verlag Enzyklopädie, 1983. Bd. 1. - 215 S.

Helbig G., Buscha J. Deutsche Grammatik. - Leipzig: VEB Verlag Enzyklopädie, 1986. - 737 S.

THE SELECTIVITY OF SENSE COLLOCATION AND POSSIBLE WAYS OF ITS STATISTICAL EXPRESSION

Vasily Ivanyuk, Victor Levitsky

S u m m a r y

The article presents the results of a statistical study of the usage of 4 semantic classes of German verbs in their different tense forms. With the help of the coefficient of contingency it has been found out that certain lexical meanings possess an ability to collocate with definite grammatical meanings. The results of the investigation have also shown that the usage of different semantic classes of verbs in this or that tense is non-arbitrary and is in conformity with certain regularities.

ЕЩЕ РАЗ О ДИФФЕРЕНЦИАЦИИ ТИПОВ АНГЛИЙСКОГО НАУЧНО-ТЕХНИЧЕСКОГО ТЕКСТА

Н.Манасян.

В предыдущей работе автора, сданной в печать в настоящую серию, вопрос о дифференциальных признаках рассматривался на основе дифференцирующих возможностеи таблиц сопряженности, категоризованными признаками в которых являются тип научно-технического текста (ТНТТ) и лингвистическии признак (ЛП).

В данной работе этот вопрос рассматривается с помощью установления значимости различия между встречаемостью ЛП в двух группах ТНТТ). (Тип текста (ТТ) в данной работе отождествляется с жанром).

Для лингвостатистического описания ТНТТ было отобрано II ТТ: монография, статья, учебник, задачник, аннотация, справочник, рекламный проспект, патент, деловое письмо, статья в профессиональнои газете и научная дискуссия.

В качестве выборочных представителей выступали отрезки текста равной длины. Величина достоверной эмпирической частоты при относительной ошибке в 33%, надежности 0.95 и длине выборки I000 словоупотреблений составила 36.

Сущностью задачи является выяснение вопроса о дифференцирующих свойствах ЛП по отношению к группе ТНТТ. Для ответа на этот вопрос обследуемые тексты разделяются на две группы по какому-либо признаку, например, на тексты с наименьшей встречаемостью некоторого ЛП и на тексты с наибольшей встречаемостью.

Пусть n_0 – это объем первой группы, а n_1 – объем второй группы, x – общее число встречаемости данной градации признака в первой группе текстов, а y – во второй. Принимаем, что проверка каждого словоупотребления на данную градацию (наугад взятая лингвистическая единица) – это независимое событие с одной и той же вероятностью успеха для данной группы текстов. В силу того, что мы имеем независимое событие с постоянной вероятностью успеха, применимо биномиальное распределение. Обозначим вероятность успеха для первой группы текстов через p_0, а для второй через p_1. p_0 и p_1 нам неизвестны, но $p_0 \approx x/n_0$, $p_1 \approx x/n_1$. (принимаем, что выборка достаточно велика).

В силу сделанных предположений случайная величина x распределена по биномиальному закону с параметрами $n_0 p_0$, а y по биномиальному закону с параметрами $n_1 p_1$. Выдвигается гипотеза H_0, что $p_0 = p_1$ и альтернативная гипотеза H_1, что $p_0 < p_1$ (Для наших целей желательно отвергнуть H_0 в пользу H_1).

Составим случайную величину $z = \frac{y}{n_1} - \frac{x}{n_0}$ (разность долей успехов).

В силу того, что n_0 и n_1 велики, применима предельная теорема Муавра-Лапласа, по которой величина z распределена по нормальному закону со средним значением и дисперсиеи D_z, где

$$D_z = \frac{p_0(1-p_0)}{n_0} + \frac{p_1(1-p_1)}{n_1} \approx$$

$$\approx \frac{1}{n_0}\frac{x}{n_0}\left(1-\frac{x}{n_0}\right) + \frac{1}{n_1}\frac{y}{n_1}\left(1-\frac{y}{n_1}\right).$$

Если верна гипотеза H_0 , то в силу того, что n_0 и n_1 велики, можно считать случайную величину $\tilde{z} = z\sqrt{D_z}$ распределенной по нормальному закону с параметрами 0,I.

По таблицам стандартного нормального закона находим $z_{крит}$ для заданного уровня значимости α и сравниваем вычисленное z с $z_{крит}$. Если $z_{выч} > z_{крит}$, то гипотеза H_0 отвергается в пользу гипотезы H_1 с уровнем значимости α.

Проверку на величину значимости различия между встречаемостью ЛП в двух группах ТНТТ прошли градации ЛП разных лингвистических уровней. Рассмотрим результаты вышеописанного анализа дифференцирующих способностеи градации ЛП "часть речи" (см. таблицу).

Тексты группировались как по минимальной и максимальной встречаемости данной градации, так и по подстилевому и жанровому признаку. В качестве частного случая группировки текстов может выступать один ТТ.

Результаты анализа показали, что ярче остальных частей речи дифференцирующая способность проявилась у части речи "существительное": в II случаях из обследованных I3 эта градация дифференцирует ТНТТ.

Далее следуют в порядке убывания следующие части речи: предлог и местоимение – I0 случаев дифференциации, артикль –

7, союз - 6, наречие и прилагательное - 4. Подобные результаты, по-видимому, связаны с номинативным характером научно-технического стиля.

Результаты анализа также дали возможность заключить, что с точки зрения морфологического состава самыми однородными из проанализированных группировок является ТТ "бюллетень" и "обсуждение доклада" (всего две части речи из восьми обладают здесь типоразличающими способностями), а самыми разнородными с этой точки зрения являются "статья" и "учебник" - их различают 7 частей речи.

Таблица

Дифференцирующая способность градации ЛП "часть речи"

I	2	3	4
Тип группировки	Уровень значимости	Наличие дифференциальных свойств градации (в порядке убывания значения ℱ)	Отсутствие дифференциальных свойств градации (в порядке убывания значения ℱ)
Минимальное/максимальное значение встречаемости градации ЛП	0.05	Глагол, местоимение, личная форма глагола, Причастие I, Причастие II, наречие, предлог, союз	
	0.10	Прилагательное	
			Местоимение, артикль
Академический подстиль (статья, монография)/ учебный подстиль (учебник, задачник)	0.05	Существительное, наречие, глагол	
	0.10	Предлог	
	2.50	Местоимение	
			Артикль, прилагательное, союз

I	2	3	4
Статья/ бюллетень	0.05	Существительное,	
	0.I0	Местоимение, предлог	
	5.00	Союз	
			Прилагательное, глагол, наречие, артикль
Статья/ деловое письмо	0.05	Местоимение, существительное, предлог	
	5.00	Глагол, союз	
			Прилагательное, артикль, наречие
Статья/ монография	0.05	Артикль, глагол	
	0.25	Глагол, существительное	
	I.00	Союз	
	5.00	Предлог	
			Наречие, прилагательное, местоимение
Статья/ задачник	0.05	Глагол, существительное, артикль	
	0.50	Предлог	
	5.00	Наречие	
			Прилагательное, союз, местоимение
Статья/ реклама	0.05	Существительное, местоимение, предлог, наречие	
	5.00	Прилагательное	
			Союз, глагол
Статья/ обсуждение доклада	0.05	Местоимение	
	2.50	Существительное	
	5.0C	Предлог, артикль	
			Союз, наречие, глагол, прилагательное

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

ONCE AGAIN ON THE DIFFERENTIATION OF ENGLISH TECHNOLOGICAL TEXTS

Narinay Manasyan

S u m m a r y

In this paper the problem of text differentiation is regarded with the help of the calculation of linguistic variable distinction significance.

It is shown that the nown shows the strongest text distinctive capacity among other parts of speech.

ПРИНЦИПЫ ФОРМАЛЬНОГО РЕШЕНИЯ ПРОБЛЕМЫ СООТНОШЕНИЯ МЕЖДУ ТЕРМИНОМ И СЛОВОМ

В.Е. Остапенко

ВВЕДЕНИЕ

Проблема соотношения между термином и словом, а также проблемы формального выделения первого из системы языка или языка той или иной науки являются достаточно сложными, противоречивыми и, как следствие, не имеющими однозначного или по крайней мере приемлемого формального решения. Как показала практическая работа по упорядочению и нормализации терминологии, основная трудность состоит в том, что рабочие определения терминов, на основе которых проводится это упорядочение, сами по себе отличаются неполнотой и нестрогостью, что в свою очередь приводит к периодической переоценке самих подходов, на основе которых формулируются эти определения.

Авторы обзорных работ, посвященных данной проблематике (см., например, Головин 1980, 1981; Лейчик 1976; Лейчик, Смирнов 1981; Табанакова 1982; Шелов, Мясников 1987; Циткина 1987), отмечают, что дискуссия ведется по следующим вопросам:

1) насколько правомерно выделение термина в качестве особой единицы языка, т.к. есть немало оснований полагать, что термины - это не особые слова, а только слова в особой функции;

2) насколько правомерно определение терминов в качестве слов в номинативной функции, т.к. существуют примеры терминирования неноминативных форм;

3) каково положение термина в знаковой системе языка и знаковой системе той или иной науки, т.е. является ли данное слово термином лексиса или термином логоса;

4) какова процедура разграничения общеупотребительной и терминологической лексики.

Нетрудно заметить, что столь явные и внешне равнозначные противоречия во взглядах на термин обусловлены отсутствием формальной, независимой от субъекта процедуры классификации лексики, которая была бы сопоставимой с системой доказательств мате-

матических теорем. При этом в отличие от обычной лингво-статистической практики не следует ограничивать роль математики чисто инструментальными или прикладными задачами, поскольку в данном случае необходима математизация нечеткого гуманитарного мышления (Шрейдер 1978). Иными словами, вышеуказанные положения (см. пп.1-3) следует рассматривать как своего рода лингвистические теоремы, которые можно доказать, опровергнуть или найти меру их истинности при помощи формальных процедур, которые представляют собой систему доказательств.

Подобная постановка задачи является встречным движением к идеям Л.Заде (Заде 1976), который предложил отказаться от высоких стандартов точности при моделировании поведения гуманистических систем, полагая, что приближенные способы рассуждений могут оказаться более созвучными сложности и неточности подобных систем, чем обычные численные методы анализа.

Попытка внести некоторую четкость в лингвистические рассуждения, т.е. попытка решить обратную задачу, и составляет сущность данной работы.

I. СУЩНОСТЬ МАТЕМАТИЗАЦИИ ГУМАНИТАРНОГО МЫШЛЕНИЯ И СОДЕРЖАНИЕ ИСХОДНЫХ ПОНЯТИЙ

Попытаемся конкретизировать существо предлагаемого подхода и содержание исходных понятий на примере частотного словаря (ЧС).

Если проанализировать лексико-статистическую структуру ЧС отраслевого текста, где приводится полный список словоформ, то можно заметить одну регулярно повторяющуюся закономерность, а именно: независимо от языка и отрасли научно-технического знания, первые ранги упорядоченного множества лексических единиц (ЛЕ) заняты служебной лексикой (приблизительно $1 \leq i \leq 15$), после чего однородность этой группы постепенно нарушается общеупотребительной и терминологической лексикой. При этом следует заметить, что интервал $16 \leq i \leq 25$ также занят в основном служебной лексикой и здесь логично говорить о приемлемой однородности группы, уровень которой может быть задан произвольно.

Если теперь весь предшествующий опыт по составлению ЧС рассматривать в качестве математического доказательства истинности суждения, то его можно предварить как четкими, так и нечеткими теоремами, справедливыми для всей последующей практики. При этом вслед за Л.Заде под теоремами будем подразумевать утверждения общего вида; если А, то В, значение истинности которых выводится из системы аксиом или же доказывается путем повторного применения общего правила, приводящего в пределах оговоренной точности к однозначному воспроизведению результата (Алимов 1980), т.е.

Теорема I.I. Если слова достаточно большого текста упорядочены по убыванию частоты их употребления, то в начале ранжированного множества ЛЕ ($I \leq i \leq I5$) образуется однородная группа служебных слов.

Теорема I.2. Если слова достаточно большого текста упорядочены по убыванию частоты их употребления, то в начале ранжированного множества ЛЕ ($I6 \leq i \leq 25$) образуется приемлемо однородная группа служебных слов.

Отсюда легко выводятся следствия, позволяющие с известной точностью решать задачи распознавания лексического класса слова, поскольку большинство реальных классов размыты по своей природе в том смысле, что переход от принадлежности элемента той или иной группе ЛЕ скорее постепенен, чем скачкообразен, т.е. для данного элемента X и класса У в большинстве случаев вопрос состоит не в том, принадлежит ли X к У, а в том, в какой степени X принадлежит или не принадлежит У (Заде 1980), т.е.:

Следствие I.I Если ЛЕ лежит в интервале $I \leq i = I5$ ЧС, то она принадлежит классу служебной лексики, а мера ее принадлежности этому классу тем больше, чем ближе эта ЛЕ расположена к началу множества.

Следствие I.2 Если ЛЕ лежит в интервале $I6 \leq i = 25$ ЧС, то она скорее всего принадлежит классу служебной лексики, т.е. в отличие от предыдущего случая принадлежность данной ЛЕ классу служебной лексики носит вероятностный характер.

Теоретическую базу остальных положений данной работы составляют соответствующим образом переработанные понятия теории

нечетких множеств (Заде 1976), теории систем (Юдин 1978) и лингво-математического моделирования (Пиотровский, Рахубо, Хажинская 1981), а также некоторые положения теории статистических группировок (Кендалл, Стьюарт 1986), т.е.

1. Будем считать размытым такое множество языковых единиц, которое представляет собой их непредсказуемую последовательность. Типичным примером такого множества можно считать ЧС отраслевого текста, т.к. в словаре этого типа основной инвентарь языковых единиц, начиная приблизительно с $i > 25$, представляет собой непредсказуемую последовательность служебной, общеупотребительной и терминологической лексики.

2. Будем называть четким такое множество языковых единиц, в котором существует четкая граница между группами слов, принадлежащих разным лексическим классам. Например, тот же ЧС отраслевого текста мог бы считаться четким множеством, если бы он представлял собой совокупность непересекающихся лексических классов.

3. Будем называть приемлемо четким такое множество, в котором существует предсказуемая по месту и составу последовательность языковых единиц, принадлежащих одному и тому же лексическому классу. Например, по отношению к служебной лексике ЧС может считаться приемлемо четким множеством, поскольку в его высокочастотной зоне (приблизительно $1 \leq i \leq 15$) всегда наблюдается однородное скопление служебных слов, которое по мере перехода в зону средних частот размывается общеупотребительной и терминологической лексикой.

4. Будем называть кластером общеупотребительной, терминологической и т.д. лексики приемлемо однородные группы слов, локализованные в определенных зонах упорядоченного множества. Типичным примером такого кластера может служить высокочастотная зона ЧС $1 \leq i \leq 20$, где в основном сосредоточена служебная лексика.

5. Будем считать приемлемо чистым или приемлемо однородным такой кластер, который содержит не менее 80% языковых единиц, принадлежащих одному и тому же лексическому классу. В принципе этот уровень задается исследователем произвольно и его можно изменить в ту или иную сторону в зависимости от тре-

бований, предъявляемых к чистоте выделяемой группы.

6. Будем называть дисперсионной моделью текста(ДМ) множество языковых единиц, упорядоченное по убыванию дисперсии их распределения.

2. РЕШЕНИЕ ПРОБЛЕМЫ СООТНОШЕНИЯ МЕЖДУ ТЕРМИНОМ И СЛОВОМ ПРИ ПОМОЩИ ДМ

В качестве экспериментального материала, на котором отрабатывалась процедура формального решения поставленной задачи, был использован французский текст по железобетону объемом 200 тыс. словоупотреблений, который был разбит на 20 сегментов. После подсчета частот словоформ в каждом сегменте и отбора знаменательных слов с абсолютной частотой $1554 \geq F \geq 22$ их употребления в тексте была сформирована совокупность из 900 ЛЕ, где термины составляли приблизительно половину (54%), а номинативные слова - около 40%.

Теперь, исходя из предположения, что распределение общеупотребительных и неноминативных слов в техническом тексте должно быть более равномерным, чем распределение терминов, что с количественной стороны это может быть выражено соответственно меньшими и большими значениями дисперсии

$$D = \sum_{k=1}^{20} \frac{(F_i - \bar{F})^2}{k-1}$$

где F_i - частота употребления ЛЕ в каком-либо из фрагментов текста,

$\bar{F}$ - средняя частота употребления ЛЕ,

k - количество фрагментов, которое в нашем случае равно 20,

и что, упорядочив ЛЕ по убыванию значений этого систематизирующего признака, можно получить однородные или приемлемо однородные лексико-грамматические группы слов, построим ДМ текста, структура которой в нашем случае оказалась следующей (см.табл.):

Таблица

СТРУКТУРА ДМ

$i_н - i_к$	А	Н	О	Т	$i_н - i_к$	А	Н	О	Т
I-25	-	I00	-	I00	30I-325	÷	÷	20	80
26-50	-	I00	-	I00	÷ ÷ ÷	÷	÷	÷	÷
5I-75	4	96	-	I00	65I-675	84	I6	÷	÷
76-I00	8	92	-	I00	676-700	88	I2	÷	÷
I0I-I25	4	96	4	96	70I-725	84	I6	÷	÷
I26-I50	20	80	I6	84	726-750	76	24	÷	÷
I5I-I75	÷	÷	I2	88	75I-775	88	I2	÷	÷
I76-200	÷	÷	36	64	776-800	80	20	92	8
20I-225	÷	÷	24	76	80I-825	96	4	80	20
226-250	÷	÷	24	76	826-850	96	4	80	20
25I-275	÷	÷	I6	84	85I-875	84	I6	80	20
276-300	÷	÷	20	80	876-900	I00	-	I00	-

Примечание:

$÷ ÷$ - зона размытости;

$i_н - i_к$ - начало и конец анализируемого интервала;

$А, Н$ - доля в % соответственно неноминативной и номинативной лексики;

$О, Т$ - доля в % соответственно общеупотребительной и терминологической лексики.

а) наблюдается очевидное тяготение терминологической и номинативной лексики к началу множества, где доля единиц этих классов составляет I00% в довольно широком интервале, с одновременным столь же отчетливо выраженным тяготением общеупотребительной и неноминативной лексики к концу множества, что вполне характеризует его в качестве совокупности ЛЕ с приемлемо четкой периферией и размытым ядром;

б) уменьшение или увеличение доли терминов в том или ином интервале однозначно связано с уменьшением или увеличением доли номинативной лексики и, наоборот, уменьшение или

увеличение доли общеупотребительной лексики в интервале связано с аналогичным изменением доли неноминативной лексики;

в) начиная приблизительно с i = 150 до i = 775 ДМ асимметрична относительно лексико-грамматических классов, т.е. приемлемая однородность кластера терминологической лексики необязательно соответствует приемлемо однородному кластеру номинативной лексики (см. интервал $151 \leq i \leq 325$) и, наоборот, приемлемая однородность кластера неноминативной лексики необязательно соответствует приемлемо однородному кластеру общеупотребительной лексики (см. интервал $651 \leq i \leq 775$), однако общая характеристика структуры ДМ (см. выше п. б) сохраняется;

г) неноминативные слова, нарушающие однородность кластера номинативной лексики в интервале $50 \leq i \leq 125$, являются адъективными и причастными формами номинативных терминов, что показывает возможность терминирования неноминативных форм слова.

Таким образом, на основе проведенного опыта по выявлению соотношения между термином и словом можно сделать следующие выводы:

1) номинативные слова в большей степени, чем любые другие формы слова употребляются в терминологической функции, поскольку ранги наибольшей принадлежности классу терминологической лексики заняты номинативными формами и, наоборот, ранги наименьшей принадлежности классу терминов заняты неноминативной лексикой;

2) терминологичность адъективных и причастных форм слова выражена менее отчетливо, чем аналогичная функция номинативных форм;

3) появляется возможность установить лингвистический рейтинг терминологической функции слова, где первое место будет принадлежать имени существительному, а второе - адъективным и причастным формам, имеющим в качестве прототипа термин в номинативной форме, в то время как рейтинг терминологичности остальных частей речи остается неизвестным ввиду недостаточности испытаний.

Таким образом, при помощи ДМ и понятия "мера принадлежно-

сти ЛЕ языковому классу" можно найти приемлемый ответ на один из спорных вопросов о соотношении между термином и словом (см. Введение п.2.).

Далее, на основе проведенного опыта можно сформулировать его результаты в виде четких и приемлемо четких теорем, справедливых для всех подобных испытаний, т.е. если воспроизводятся все условия построения ДМ, т.е.:

Теорема 2.1. Если слова упорядочены по убыванию дисперсии их распределения в тексте или их выборочной совокупности, то в начале ранжированного множества образуется однородный кластер терминологической лексики $I \leq i \leq 50$.

Теорема 2.2. Если слова упорядочены по убыванию дисперсии их распределения в тексте, то в начале ранжированного множества образуется кластер номинативной лексики ($I \leq i \leq 25$).

Теорема 2.3. Если слова упорядочены по убыванию дисперсии их распределения в тексте, то в начале ранжированного множества образуется приемлемо однородный кластер терминологической лексики $51 \leq i \leq 150$.

Теорема 2.4. Если слова упорядочены по убыванию дисперсии их распределения в тексте, то в начале ранжированного множества образуется приемлемо однородный кластер номинативной лексики $26 \leq i \leq 125$.

Теорема 2.5. Если слова упорядочены по убыванию дисперсии их распределения в тексте, то в конце ранжированного множества образуется однородный кластер неноминативной и нетерминологической лексики объемом до 25 ЛЕ.

Теорема 2.6. Если слова упорядочены по убыванию дисперсии их распределения в тексте, то в конце ранжированного множества образуется приемлемо однородный кластер неноминативной и нетерминологической лексики объемом до 125 ЛЕ.

Аналогично вышесказанному (см. п.1, следствия 1-2), можно сформулировать четкие и приемлемо четкие следствия из предыдущих теорем, позволяющие решать задачи распознавания лексико-грамматического класса слова при помощи данной ДМ, т.е.

Следствие 2.1. Если ЛЕ попадает в интервал $I \leq i \leq 50$ ДМ, то она принадлежит классу терминологической лексики, а мера ее принадлежности этому классу тем выше, чем ближе она распо-

ложена к началу множества.

. . . . . .

Следствие 2.6. Если ЛЕ попадает в конечный интервал ДМ, исключая последние 25 ЛЕ, то вероятнее всего она принадлежит классу неноминативной лексики.

ЗАКЛЮЧЕНИЕ

Результаты проведенного опыта по формальной классификации ЛЕ и математизации гуманитарного мышления показывает, что существует возможность математического решения проблемы соотношения между термином и словом, т.е. спорные вопросы теоретического языкознания могут решаться не умозрительно, а опытным путем, в частности, при помощи ДМ.

При этом необходимо отметить, что с увеличением объема текста до 300 тыс., 400 тыс. и т.д. словоупотреблений станет более четкой структура ДМ, т.е. увеличатся интервалы 100%-ной принадлежности классу терминов классу номинативной лексики и т.д., а также увеличатся соответствующие приемлемо однородные кластеры на периферии ранжированного множества. Это означает, что одновременно повысится четкость выше сформулированных теорем и следствий.

В данной работе была предпринята попытка математического решения лишь одного из трудных вопросов теоретического языкознания, однако есть основания полагать, что существует аналогичное решение и двух других вопросов (см. Введение). Очевидно для этого необходимы дополнительные модели и их сопряжение в систему, обладающую синергетическими свойствами.

Л И Т Е Р А Т У Р А

Алимов Ю.И. Альтернатива методу математической статистики. - М.: Знание, 1980. 64.с

Головин Б.Н. Термин и слово. //Термин и слово. - Горький: Изд-во ГГУ, 1980. - С.3-12.

Головин Б.Н. Типы терминосистем и основания их различения. //Термин и слово. Межвузовский сборник. - Горький: Изд-во ГГУ, 1981. - С.3-10.

Заде Л.А. Понятие лингвистической переменной и его применение к принятию приближенных решений. - М.: Мир, 1976. - С.5-146.

Заде Л.А. Размытые множества и их применение в распознавании образов и кластер-анализе. //Классификация и кластер. - М.: Мир, 1980. - С.208-247.

Кендалл М.Дж., Стьюарт А. Многомерный статистический анализ и временные ряды. - М.: Наука, 1976. - С.437-468.

Лейчик В.М. Термины и терминосистемы - пограничная область между естественным и искусственным в языке. //Вопросы терминологии и лингвистической статистики. - Воронеж: Изд-во Воронежского университета, 1976. - С.3-11.

Лейчик В.В., Смирнов И.П. Современное состояние и тенденции дальнейшего развития терминологической работы в СССР. //НТИ ВИНИТИ. Сер.2. - № 3. - С.4-7.

Пиотровский Р.Г., Рахубо Н.П., Хажинская М.С. Системное исследование лексики научного текста. - Кишинев: Штиинца, 1981. - С.5-8.

Табанакова В.Д. Понятие научно-технического термина и требования к его определению. //Термин и слово. - Горький: Изд-во ГГУ, 1982. - С.24-28.

Циткина Ф.А. Интернациональный термин в статистике и динамике: перспективы упорядочения с точки зрения сопоставительного терминоведения. //НТИ ВИНИТИ. Сер.2. - 1987. - № 5. - С.5-9.

Шелов С.Д., Мясников А.Г. Логико-семантическая структура терминологии и ее формальные свойства. //НТИ ВИНИТИ. Сер.2. - 1987. - № 3. - С.6-10.

Шрейдер Ю.А. Гуманитаризация знания и управление информационной средой.// Вестник АН СССР №9. -М.: Изд-во АН СССР, 1978. -С.24-39.

Юдин Э.Г. Системный подход и принцип деятельности. -М.: Наука, 1978. -С.96-245.

PRINCIPES DE SOLUTION FORMELLE DU PROBLEME DE CORRELATION ENTRE UN TERME ET UN MOT

Vladimir E. Ostapenko

R é s u m é

L'article traite les perspectives de mathématisation de mentalité linguistique visant à la solution du problème de correlation entre un terme et un mot.

On considère les contradictions des conceptions linguistiques concernant la correlation précitée et on cherche la solution du problème sous forme de théorèmes qui peuvent être démontrées à partir d'un modèle mathématique.

On dresse le principe de modélisation statistique qui peut être présenté sous forme d'un ensemble de mots soumis en ordonnance descendante de la dispersion de leur distribution dans le texte.

On considère aussi les perspectives de l'extraction formelle de termes à partir de la structure statistique du modèle dressé ainsi que celles de reconnaissance d'une classe linguistique du mot.

О НЕКОТОРЫХ СОДЕРЖАТЕЛЬНЫХ ХАРАКТЕРИСТИКАХ СТИЛЯ

С.О. Савчук

Основной проблемой, возникающей при использовании количественных методов в стилистических исследованиях, является проблема выбора параметров для количественной характеризации стиля. В подавляющем большинстве случаев в качестве таких параметров используются поверхностно-языковые признаки, и это обстоятельство вызывает справедливые замечания о том, что такой выбор, безусловно, "значительно упрощает операции измерения, но не исключено, что наряду с этим упрощает и суть дела" /Богданов, 1989, с. 160/. Исходя из этого, большую актуальность приобретает поиск связи между глубинными, содержательными характеристиками стиля и их поверхностным, формально-языковым выражением, поддающимся количественному анализу. В настоящей статье обосновывается один из вариантов такого стилистического анализа, при котором в качестве стилистических параметров используются показатели, непосредственно отражающие действие основных групп стилеобразующих факторов, и в силу этого "улавливающие" существенные, содержательные характеристики стиля.

Избранный подход базируется на двух исходных предпосылках. В качестве первой предпосылки принято такое определение стиля, согласно которому стиль понимается не как тот или иной набор каких-либо языковых средств, но как тот или иной способ языкового оформления коммуникативного намерения говорящего. Данное определение имеет два аспекта. С одной стороны, в процессуальном аспекте, понимаемый как способ оформления коммуникативного намерения, рассматривается как последовательность определенных коммуникативных действий говорящего в процессе реализации замысла. С другой стороны, в результативном аспекте, предполагается понимание стиля и как формы текста, то есть как способа организации содержания текста с помощью языковых средств /поскольку создание текста является конечной целью и результатом речевой деятельности говорящего/.

Вторым теоретическим допущением является идея о том, что выбор говорящим стратегии стилистического действия не стихи-

ен, но детерминирован влиянием стилеобразующих факторов. В качестве основных, базовых стилеобразующих факторов, к которым сводится все многообразие их частных проявлений, в работе принята выдвинутая М.М. Бахтиным триада: /I/ отношение говорящего к предмету речи, /2/ отношение говорящего к адресату и /3/ отношение говорящего к чужим речам о выбранном предмете. Эти три вида отношений Бахтин считал основными конститутивными моментами высказывания, теми факторами, под влиянием которых происходит языковое оформление смысла высказывания, формирование его стиля.

Для того чтобы обнаружить конкретное языковое влияние этих базовых факторов на стиль, что равносильно выявлению сущностных характеристик стиля, необходим специальный инструмент, отличный от распространенных приемов стилистического анализа и способный фиксировать проявление действия названных факторов[+]. Для этой цели был выбран аппарат лингвистического анализа, предложенный в работах Гоготишвили 1984 и 1985, что вызвано, во-первых, тем, что он основан на учете всех трех названных выше стилеобразующих факторов вместе и, во-вторых, тем, что он опирается на традицию М.М. Бахтина. При этом общелингвистическое понимание проблемы, свойственное этому автору, сменилось в настоящей работе ее непосредственно стилистической интерпретацией. Суть предложенной методики анализа текста состоит в том, что процесс порождения речи мыслится здесь как последовательность поочередно сменяемых языковых действий, которые определяются внешними по отношению к самой речи факторами. По терминологии Л.А. Гоготишвили, "общая смысловая позиция говорящего"/ то есть условно зафиксированный полный объем смысла данного текста/ воплощается в речи с помощью сложной системы частных источников смысла. Частный /по отношению к общей смысловой позиции/ источник смысла /ЧИС/ это некоторая зафиксированная в тексте смысловая или пространственно-временная позиция, с которой говорящий подает оче-

[+] Попытка создания такого инструмента для анализа текста и в типологических целях можно найти в целом ряде работ. См., например, Вежбицка, 1978 и 1982; Золотова, 1979; Дейк, 1978; Кожевникова, 1985, Нунэн, 1982; Славгородская, 1981; Успенский, 1970; Чейф, 1982 и др.

редной содержательный фрагмент высказывания в соответствующей языковой форме /Гоготишвили, 1985, с. 155/. Выделяются три вида частных источников смысла /ЧИС/: речевой центр /РЦ/, точка зрения /ТЗ/ и фокус внимания /ФВ/. Каждый вид ЧИС можно поставить в соответствие одному из трех упомянутых выше базовых стилеобразующих факторов: речевой центр может быть интерпретирован как отражение влияния фактора чужой речи, точка зрения - фактора адресата, а фокус внимания, соответственно, фиксирует действие третьего фактора - отношения говорящего к предмету речи.

Центральным моментом данной методики являются не сами ЧИС, но факт их последовательной, частой и неизбежной для любой речи смены/процессуальный аспект речи/. Эта способность к смене на протяжении строящегося высказывания является общим свойством всех ЧИС. Другой общей особенностью всех видов частных источников смысла является то, что смены ЧИС не связаны жестко с какой-то определенной языковой формой /как, например, грамматическое значение, выражающееся в строго определенном классе форм/, а, напротив, могут осуществляться различными языковыми способами. Выявление различных способов языкового выражения смен ЧИС представляется необходимой ступенью, предшествующей использованию ЧИС в методике стилистического анализа, и служит установлению непосредственной связи между глубинными, латентными характеристиками стиля и их поверхностным формально-языковым выражением.

Остановимся подробнее на особенностях каждого вида ЧИС.

Речевой центр /РЦ/ - это смысловая позиция, принадлежащая либо самому автору, либо конкретному оппоненту, либо различным видам социально-обобщенных точек зрения. Так, например, в конструкциях с прямой речью "слова автора" подаются из авторского речевого центра, а "прямая речь" принадлежит чужому речевому центру. В качестве примера более сложного взаимодействия различных РЦ рассмотрим отрывок из статьи о молодом учителе музыки из провинциального городка, который, случайно попав на международный конкурс пианистов, завоевал там главную награду. Рассказ о казалось бы счастливой судьбе музыканта наводит автора статьи на размышления о недостатках, вернее, о полном отсутствии системы поиска и воспита-

ния отечественных талантов.

Поле поиска /I/ своих Платонов и быстрых разумом Невтонов /2/ сузилось до границ Московской кольцевой автодороги. Утешения по поводу того, как /3/ люди добрые /4/ все ж таки /5/ паренька приметили /6/ - они для /7/ шахты угольной /8/, к тому же довоенной. /Л. Парфенов. Не бойтесь находить таланты//"Огонек", 1988, № 24/.

В данном случае выделенные фрагменты взяты автором из чужих контекстов /чужих РЦ/: первый - из оды М.В. Ломоносова, второй - из некогда популярной песни. При этом использование чужих слов не является для автора самоцелью, но производится для столкновения в одном контексте различных смысловых позиций, которые стоят за соответствующими цитатами. Так, строка из оды Ломоносова заставляет вспомнить фигуру самого основателя Академии и университета, о котором существует мнение, что он отдавал много сил воспитанию талантов; за словами из песни скрывается расхожее мнение о том, что талант, дескать, всегда пробьется, всегда себе дорогу найдет и т.д., то есть позиция, с которой автор категорически не согласен и против которой возражает в своей статье. Таким образом, выражая свою мысль не прямо, а используя различные РЦ, автор обогащает смысл своего высказывания дополнительными смысловыми обертонами. Включение говорящим "чужих" слов и смыслов в свое высказывание производится с помощью механизма смен речевых центров. Границы смен РЦ отмечены цифрами: /I/, /3/, /5/, /7/ обозначают смены РЦ автора /РЦА/ на чужие РЦ - Ломоносова /I/ и популярной песни /3/, /5/, /7/ цифры /2/, /4/, /6/, /8/ отмечают обратные смены с соответствующих чужих речевых центров на РЦ автора. Смены речевых центров в данном примере производятся не с помощью типичных для передачи чужой речи способов прямой и косвенной речи, напротив, границы этих смен никак формально-синтаксически не отмечены и проходят внутри предложения и даже внутри словосочетания /способ "рассеянной чужой речи" по М.М. Бахтину/Волошинов, 1929/. В результате различные РЦ совмещаются либо диссонируют в одном слове /возникает т.н. "эффект двуголосого слова" М.М. Бахтина/.

Точка зрения /ТЗ/ - это смысловая позиция, которую зани-

мают в акте коммуникации ее основные участники - говорящий и слушающий. Позиция говорящего обозначается как ТЗ-Я, слушающего - как ТЗ-ТЫ. Поскольку речь идет, разумеется, не о реальном слушающем, а о некоторой концепции адресата, конструируемой говорящим, то ТЗ - ТЫ представляет собой такую, основанную на концепции апперцептивного фона слушающего, позицию, на которую временно становится говорящий, чтобы с точки зрения адресата оценить строящееся высказывание и либо устранить нежелательные моменты возможного непонимания, уточнить свою позицию, либо, предвосхитив возможное непонимание, вовсе избежать его. Приведем пример:

...На этом пути пока сделано очень мало.../I/ Достаточно сказать /2/, что у нас вовсе нет словаря русского языка, ...нет хорошего этимологического словаря /3/ /Преображенский, как известно, остался незаконченным/ /4/../5/И говорить нечего /6/, что почти нет хороших лингвистических разборов литературных произведений.../7/ Что же делать? /8/ Содействовать появлению соответственных трудов... /Л.В. Щерба. Безграмотность и ее причины // Избранные работы по русскому языку. - М., 1957. С. 60-61/.

В данном примере цифры отмечают границы смен точек зрения ТЗ говорящего и ТЗ слушающего, с помощью которых осуществляется временный переход автора на позицию слушающего, своеобразный диалог его с адресатом, "реплики" которого /в тексте они подчеркнуты/ также включаются в готовое высказывание: /I/ - смена ТЗ-Я на ТЗ-ТЫ, /2/ - обратная смена ТЗ-ТЫ на ТЗ-Я, /3/ - ТЗ-Я → ТЗ-ТЫ, /4/ - ТЗ-ТЫ → ТЗ-Я, /5/ - ТЗ-Я → ТЗ-ТЫ и т.д. Языковые способы смен точек зрения разнообразны: /I/ - /2/ и /5/ - /6/ - метатекстовые включения; /7/ - /8/ - риторический вопрос, который является как бы воспроизведением фрагмента внутренней речи слушающего; /3/ - /4/ - вставная конструкция, которая представляет собой предвосхищенный ответ на возможный вопрос слушающего: "А Преображенский?"

Фокус внимания/ФВ / - это тематическая позиция, в которую помещается наиболее важный в смысловом отношении компонент данного содержательного фрагмента высказывания, становящийся на время смысловым центром, вокруг которого группируются все остальные элементы содержания. На синтаксическом уровне уста-

новление фокуса внимания выражается в том, что выделяемый элемент содержания либо сам занимает определенную синтаксическую позицию, либо отмечается с помощью специальных языковых средств. "Важность" помещаемого в фокус внимания элемента смысла определяется не только коммуникативным намерением говорящего, но и воздействием фактора адресата, поскольку учет апперцептивного фона слушающего "подсказывает" говорящему, что сейчас необходимо выдвинуть на первый план, о чем, напротив, целесообразней умолчать или упомянуть вскользь и т.д. Смены ФВ /который, в отличие от предыдущих видов ЧИС, существует лишь в одной ипостаси и сменяется на самого себя/ определяют тематическое развертывание высказывания: производя смены ФВ, говорящий сосредоточивает внимание слушающего на нужных элементах содержания и прокладывает маршрут изложения информации от центра к центру, не позволяя отклоняться в сторону. Рассмотрим пример.

Женщина /I/. впустившая меня в квартиру, внесла блюдечко /2/, на котором лежала одинокая вареная морковка /3/, неаккуратно очищенная и уже немного подсохлая. /А. Найман. Рассказы о Анне Ахматовой//"Новый мир", 1989, № I. С. 161/.

Данный пример демонстрирует, как автор управляет вниманием слушающего, последовательно фиксируя его на различных объектах. Сначала фокус внимания устанавливается на одном из участников описываемой ситуации - "женщине" /I/. Делается это способом подлежащего - придавая функцию подлежащего элементу, попавшему в фокус внимания; далее, вводя определительное придаточное к слову "блюдечко" /2/, говорящий перемещает взгляд слушающего на этот элемент содержания; наконец, с помощью причастного оборота, - к содержимому блюдечка /3/. Отметим разнообразие использованных способов смен ФВ, представленных в рассмотренном примере.

Итак, суммируя сказанное, можно констатировать: I/ частные источники смысла (речевой центр, точка зрения, фокус внимания) рассматриваются в предлагаемой методике в качестве инструмента для фиксации действия стилеобразующих факторов /отношения говорящего к предмету речи, к адресату и к чужим речам о предмете/; 2/ процесс попеременной смены ЧИС трактуется как проявление существенных сторон процесса стилеобразова-

ния; 3/ языковые формы, в которых осуществляются смены ЧИС, рассматриваются как важнейшие характеристики стиля, понимаемого как способ языкового оформления смысла.

Таким образом, смены частных источников смысла, с помощью которых говорящий развертывает свое содержание, и языковое выражение этих смен, представют как две грани единого процесса образования стиля: глубинной, смысловой стороны этого процесса /ср. понятие "формы содержания" М.М. Бахтина/ и его внешнего речевого аспекта.

Опираясь на эти основания, можно выдвинуть гипотезу о том, что для каждого стиля характерен определенный набор и соотношение различных видов ЧИС, определенные способы языкового выражения этих смен, то есть перечисленные характеристики могут быть использованы для описания типологических стилистических различий.

Проверка данной гипотезы осуществлялась на материале 65 текстов различной стилистической и жанровой принадлежности общим объемом около 20 тыс. словоупотреблений. Часть текстов была подвергнута выборочному анализу, другая часть - 35 крупных отрывков объемом 250-400 словоупотреблений каждый /общий объем выборки 12 тыс. словоупотреблений/ - составили основной корпус текстов, подвергшихся сплошному многоаспектному обследованию.

Обследуемый корпус текстов отбирался с учетом ряда требований: его основу составили "образцовые" тексты - отрывки из произведений известных ученых и писателей, а также популярных публицистов; анализируемые тексты в большинстве своем отражают современное состояние русского языка; при их отборе учитывалось существующее жанровое разнообразие внутри научного, художественного и публицистического стилей.

Отобранный таким образом материал подвергся комплексному анализу, включающему три этапа: 1/ стилистический анализ текста, опирающийся на методику, предложенную в работе Гоготишвили 1984; 2/ обработку количественных данных, полученных с помощью введения ряда стилистических параметров текста; 3/ теоретический анализ данных количественного обследования, типологическую интерпретацию.

Методика стилистического анализа основана на выявлении в

тексте смен частных источников смысла, осуществляемых говорящим с помощью различных языковых способов. Текст, в котором выявлены смены ЧИС, представляется расчлененным на отрезки разной длины, границы которых /являющиеся границами смен ЧИС/ по большей части не совпадают с границами формально-синтаксического членения /ср. у Бахтина о том, что при более глубоком изучении высказывания обнаружится, что оно изнутри все "изборождено как бы далекими и еле слышными отзвуками смен речевых субъектов и диалогическими обертонами, до предела ослабленными границами высказывания"/ /Бахтин, 1979, с. 273/.

Результаты анализа по сменам ЧИС всех 35 текстов в дальнейшем переводятся на язык стилистических параметров, с помощью которых проведено комплексное обследование выбранного материала. В качестве таких параметров были использованы текстовые и стилистические показатели.

I. Общее количество словоупотреблений в тексте, включая служебные слова и элементы других коммуникативных систем /иностранные слова, цифровые обозначения, математические символы, формулы и т.д./.

2. Общее количество смен частных источников смысла в тексте.

3. Среднее количество смен ЧИС на единицу объема текста /I00 словоупотреблений/, которое вычисляется по формуле:

$$\frac{\text{число смен ЧИС}}{\text{число словоупотреблений в тексте}} \times 100$$

4. Количество смен каждого вида ЧИС /РЦ, ТЗ и ФВ/ в отдельности.

5. Соотношение различных видов ЧИС между собой; вычисляется в процентах, исходя из того, что общее количество смен ЧИС составляет I00%.

6. Общая частота смен ЧИС, или показатель того, через сколько слов в среднем происходит смена какого-либо ЧИС; представляет собой отношение объема текста к общему числу смен ЧИС.

7. Частота смен каждого вида ЧИС в отдельности; представляет собой отношение объема текста к числу смен конкретного вида ЧИС.

8. Соотношение способов языкового выражения смен ЧИС. Обследование текста с помощью этого параметра состоит в выяснении того, какие языковые способы участвуют в сменах ЧИС и сколько раз встречается каждый из использованных способов, что позволяет выявить: а/ способы языкового выражения смен ЧИС, преобладающие в данном стиле; б/ способы, регулярно встречающиеся во всех стилях и, следовательно, безразличные к стилистическому варьированию; в/ способы, характерные для какого-то одного стиля и не встречающиеся в других стилях, и, следовательно, находящиеся в зависимости от стилистических различий; г/ сочетание языкового выражения смен РЦ, ТЗ и ФВ, характерное и типичное для каждого стиля.

Количественное обследование всего корпуса текстов с помощью введенных показателей подтвердило сформулированное ранее предположение относительно реальных диагностических возможностей стилистического анализа с помощью ЧИС, согласно которому каждый стиль - научный, публицистический и художественный - характеризуется особой, свойственной только ему, совокупностью различных форм взаимодействия частных источников смысла.

Конкретные типологические различия между научным, публицистическим и художественным стилями, выявленные по ряду основных параметров, представлены в таблице.

Комментарий к таблице.

I. Первая - количественная характеристика /2 и 3 клетки таблицы/ показывает, что частота смен частных источников смысла /так же, как и среднее количество смен на единицу объема текста/ находится в зависимости от стиля/ и, следовательно, может рассматриваться как одна из стилистических характеристик/: наибольшим числом смен ЧИС /и их большей частотой/ характеризуется публицистический стиль, наименьшие значения этого признака характерны для научного стиля. Этот факт, по всей видимости, связан с широкой распространенностью в научном стиле многокомпонентных понятийных словосочетаний, функционирующих как одно слово, что приводит к расширению номинального объема текста, но не к росту количества смен ЧИС, имеющих отношение к его реальному смысловому объему.

Параметры		Научный стиль	Публицистический стиль	Художественный стиль
Общий объем текстового материала (количество словоупотреблений)		3988	4096	3947
Среднее количество смен ЧИС на ед-цу объема текста (100 словоупотреблений)		30,4	37,0	32,5
Частота смен ЧИС		3,3	2,7	3,0
Соотношение смен ЧИС:				
смены РЦ		18%	35,6%	24,3%
смены ТЗ		36,2%	23,9%	16,2%
смены ФВ		45,8%	40,5%	59,5%
Частота смен ЧИС:				
РЦ		18,3	7,6	12,7
ТЗ		9,1	11,3	18,9
ФВ		7,2	6,7	5,2
Способы языкового выражения смен ЧИС:	РЦ	предметно-аналитические, синтаксически-ненаблюдаемые (тематическая речь)	словесно-аналитические, синтаксически - наблюдаемые и ненаблюдаемые (50%) (рассеянная чужая речь)	словесно-аналитические, синтаксически - ненаблюдаемые (несобственно прямая речь)
	ТЗ	комментирующие, обязательные (метатекст)	комментирующие, обязательные и необязательные	актуализирующие, необязательные
	ФВ	общеязыковые	общеязыковые, стилистические приемы - 3%	общеязыковые, стилистич. приемы - 13%

2. Второй характеристикой, разграничивающей научный, публицистический и художественный стили, является соотношение различных видов ЧИС между собой в пределах одного стиля /4 клетка таблицы/. Для научного стиля характерно резкое преобладание смен ТЗ над сменами РЦ, что свидетельствует о меньшей ориентации автора научного текста на различные смысловые оттенки чужих позиций, которые даются здесь чаще всего в своей самой общей, почти номинальной форме /заметим, что в естественнонаучных текстах доля смен РЦ ниже, а в гуманитарнонаучных- выше среднестатистической величины/.

Напротив, преобладание в научном стиле смен ТЗ, которые отражают отношение говорящего к адресату, говорит о том, что автор научного текста значительную долю своей коммуникативной энергии затрачивает на установление контакта с адресатом. Активное взаимодействие со слушающим необходимо автору научного текста для того, чтобы представить новое научное содержание в понятной и удобной для восприятия форме, что невозможно без точного знания апперцептивного фона адресата, позволяющего говорящему выбрать оптимальную стратегию доказательства, предвидеть возможные вопросы слушающего и заложить в свой текст ответы на них и т.д.

Публицистический стиль характеризуется значительным преобладанием смен РЦ над сменами ТЗ, что, по-видимому, составляет его существенную особенность: само коммуникативное задание публицистического стиля предполагает вовлечение дифференцированных смысловых позиций - различных мнений, которые могут противопоставляться, опровергаться, подтверждаться и т.д. Использование чужих смысловых позиций в публицистическом стиле есть способ формирования авторской позиции: для публицистического стиля обычным является одновременное участие в оформлении смысла пяти-шести различных РЦ /в научном стиле эта цифра, как правило, не превышает 2-3/; кроме того, авторский голос, вступая во взаимодействие с любым из чужих РЦ, может создавать различные виды совмещенных речевых центров. С другой стороны, столкновение различных чужих позиций в публицистическом стиле служит особым стилистическим приемом убеждения слушающего /за счет чего и сокращается использование смен ТЗ/: если научный стиль как бы продвигает вперед внутреннюю логику

слушающего – и потому здесь часты смены ТЗ, то публицистический стиль "убеждает" не столько логическим развертыванием частной замкнутой темы, сколько широким эмоционально разветвленным /аксиологическим/ убеждением. Разворачивая перед читателем /с помощью смен РЦ/ открытый диалог чужих мнений о предмете речи, автор старается подвести его к признанию того, что авторская позиция более предпочтительна.

Что касается художественного стиля, то суждения относительно его особенностей имеют сугубо предварительный характер, поскольку предложенная методика анализа в применении к художественному стилю нуждается в многочисленных дополнениях и уточнениях. Такая необходимость объясняется тем, что хотя процессы смены частных источников смысла происходят и в художественном стиле /что показал и анализ соответствующих текстов/, однако сам характер ЧИС, как и характер влияния соответствующих факторов, в художественном стиле принципиально иной. Тем не менее, полученные результаты анализа художественных текстов с помощью предложенных параметров, пусть и не отражающих существо художественного стиля, интересны тем, что показывают достаточно отчетливое противопоставление художественного стиля двум другим. Так, например, художественный стиль характеризуется резким преобладанием смен фокуса внимания над всеми другими видами смен, а также большей частотой смен ФВ /см. 5 клетку таблицы/ по сравнению с другими стилями. Этому факту дается следующая интерпретация. Будучи наиболее разработанным и наиболее "сделанным" стилем, в котором конденсируются и находят свое концентрированное проявление все основные выразительные потенции языка, художественное слово усиливает в себе и глубинные общеязыковые семантические приемы, в том числе и смены ФВ. Если в менее "искусных" стилях смена ФВ как бы требует большей подготовки и текст потому растягивается, то художественный стиль демонстрирует способность к более динамичному протеканию смен ФВ, что, несомненно, насыщает смысловое пространство текста при его даже небольшом объеме.

3. Третьей характеристикой, выявляющей различия между стилями, является частота смен отдельных видов ЧИС /см. 5 клетку таблицы/. Числовые данные, показывающие, с какой час-

тотой /т.е. через сколько слов/ в тексте происходят смены РЦ, ТЗ и ФВ, в целом воспроизводят описанные выше соотношения смен разных видов ЧИС, однако показывают эти соотношения с новой стороны.

4. Четвертым критерием, "улавливающим" стилистические различия, является способ языкового выражения смен ЧИС /см. 6 клетку таблицы/. Анализ материала по данному критерию показывает, что каждый из трех стилей характеризуется а/ определенным набором языковых способов смен ЧИС; б/ определенным их сочетанием; в/ каждому из трех стилей присущ "ведущий" способ смен РЦ, ТЗ и ФВ. В частности, таким "предпочтительным" способом смен РЦ в научном стиле оказывается способ тематической речи, в публицистическом стиле - способ рассеянной чужой речи, художественный стиль "предпочитает" способ несобственно-прямой речи.

Следует отметить, что хотя конкретные результаты проведенного анализа имеют предварительный "рабочий" характер, однако и уже полученных количественных и типологических данных, вероятно, достаточно для подтверждения высокого потенциала методики стилистического анализа по сменам ЧИС. Дальнейшее применение этой методики предполагает как углубление и уточнение теоретических положений, лежащих в ее основе, так и расширение объема анализируемого материала, что, в свою очередь, приведет к постановке новых практических проблем и к развитию теоретических представлений о самом феномене стилистического многообразия речи.

ЛИТЕРАТУРА

Бахтин М.М. Проблема речевых жанров // Эстетика словесного творчества. М., 1979. - С. 237-280.

Богданов В.В. Основы стилеметрии. - Рец. на кн.: Г.Я. Мартыненко. Основы стилеметрии. Л., 1988 // Учен. Зап ТГУ. Вып. 872. Тарту, 1989. - С. 157-161.

Вежбицка А. Метатекст в тексте // НЗЛ. Вып. 8. М., 1978. - С. 402-424.

Вежбицка А. Дескрипция или цитация // НЗЛ. Вып. 13. М., 1982. - С. 237-262.

Волошинов В.Н. Марксизм и философия языка. Л., 1929.

Гоготишвили Л.А. Опыт построения теории употребления языка (на основе общефилологической концепции М.М. Бахтина). Дисс. на соиск. степ. канд. филол. наук. М., 1984.

Гоготишвили Л.А. Хронотопический аспект смысла высказывания // Речевое общение: цели, мотивы, средства. М., 1985.- С. 150-171.

Дейк Т.А. ван Вопросы прагматики текста // НЗЛ. Вып. 8. М., 1978. - С. 259-336.

Золотова Г.А. Роль ремы в организации и типологии текста // Синтаксис текста. М., 1979.

Кожевникова К. О смысловом строении спонтанной устной речи // НЗЛ. Вып. 15. М., 1985. - С. 512-523.

Нунэн М. О подлежащих и топиках // НЗЛ. Вып. 11. 1982. - С. 193-235.

Славгородская Л.В. О функции адресата в научной прозе // Лингвистические особенности научного текста. М., 1981. - С. 93-103.

Успенский Б.А. Поэтика композиции. М., 1970.

Чейф У. Данное, контрастивность, определенность, подлежащее, топики и точка зрения // НЗЛ. Вып. 11. М., 1982. - С. 277-316.

ON SOME SUBSTANTIAL CHARACTERISTICS OF STYLE

Svetlana Savchuk

S u m m a r y

The paper deals with a kind of stylistic analysis based on a number of substantial characteristics of style. These characteristics (parameters) are assumed to be closely to M. Bakhtin, are: (1) the speaker's reference to the subject of speech, (2) to the addresser and (3) to another's statements of the same subject ("chuzaya rech").

With the help of the offered method a corpus of texts belonging to three main styles (scientific, rhetorical and fictional) was analysed; the results of the analysis and their typological interpretation are adduced in the paper.

ИДЕИ ГЕРМЕНЕВТИКИ В ПРИКЛАДНОЙ ЛИНГВИСТИКЕ

С.В.Чебанов, Г.Я.Мартыненко

Настоящая статья посвящена рассмотрению того, как развитие современной прикладной лингвистики приводит к возникновению идей, сходных с идеями герменевтики. В настоящее время можно говорить не только о "герменевтизации" прикладной лингвистики (для уточнения можно было бы привязать эту линию развития о языке к александрийской, символико-аллегорической герменевтической традиции), но и о герменевтизации филологии, являющейся, однако, совершенно другим процессом, идущем почти в противоположном направлении и явно тяготеющим к антиохийской (исторической) школе герменевтики.

В предыдущей публикации (Чебанов, Мартыненко, 1990) авторами было обосновано различение основных подходов к языку. В ней же речь шла и об отношениях между этими подходами. В этом аспекте данная статья может рассматриваться как специальный анализ типологических отношений между герменевтикой и прикладной лингвистикой.

Прикладная лингвистика как определенный подход к языку характеризуется крайним разнообразием направлений, методов и решаемых задач (Богданов, Бондарко и др., 1987), которые слабо связаны друг с другом как на концептуальном, так и на операциональном уровне. Это не способствует тому, чтобы говорить о единой концепции языка в прикладной лингвистике. Речь, скорее, может идти о втягивании в сферу работы определенного массива разнородных идей и фактов, совокупность которых и позволяет говорить о качественно новом состоянии лингвистического знания. Как и другие концепции языка, прикладная лингвистика не имеет четкого места в истории лингвистических учений, поскольку многие прикладные задачи, такие, как перевод, обучение языку, дешифровка, редактирование и т.п. существуют очень давно. Ситуация стала более определенной только после того, как упомянутые прикладные задачи стали получать, начиная с середины XX века, систематическое научное обоснование, отправной точкой которого явились концепции семиологии.

Семиология – та концепция языка, которая легла в основу структурной типологии языков, процедур описания естественных

языков, точных методов дешифровки и т.п. В эти разработки вовлекался все более широкий и разнообразный материал, далеко выходящий за рамки того, что раньше интересовало языковедов, в результате сложилась самостоятельная исследовательская область, которая стала обозначаться как прикладная лингвистика. В этом своем качестве прикладная лингвистика может рассматриваться как деятельностный коррелят прагмалингвистики, хотя она и сложилась значительно позднее последней. Кроме того, следует иметь в виду, что прагмалингвистика ориентирована прежде всего на прагматику антропосферы (Сусов, 1983), в то время как прикладная лингвистика ориентирована на техносферу, а иногда требует учета особенностей зоо- и даже биосферы (Прибрам, 1975; Глезер, 1985). Сюда же примыкают и проблемы биосемиотики (Шаров, 1990). Важным при этом является и то, что существенной чертой организации как техносферы, так и биосферы является субстратная природа экспонентов соответствующих семиотических систем.

Для прикладной лингвистики существенно семиологическое представление о языке как системе и структуре, что открывает возможность для формально-логического и математического описания языка (Хомский, 1962). Такой подход интерпретирует языковую деятельность как особый тип исчисления, вводит естественный язык в ряд языков другой природы (искусственных, машинных). Это является базой для создания лингвистических автоматов, реализуемых в конечном счете на компьютерной основе.

Исходным пунктом разработки лингвистического автомата является представление о языке как системе чистых отношений между знаками. В последних существенно наличие эксплицитно представленного экспонента, что создает основу для их автоматической обработки. В соответствии с концепцией структурной лингвистики семантические моменты появляются в результате анализа эксплицитно представленных синтаксических структур. Такая постановка вопроса вывела на передний край лингвистических исследований план выражения, вокруг которого группируются два больших направления прикладной лингвистики.

Первое из них занимается исследованием письменного текста - печатного и рукописного.

При анализе печатного текста решается широкий круг прикладных задач. Во-первых, это оптимизация начертания графем

с целью обеспечения эффективности их зрительного и машинного распознавания. Решение этой задачи предполагает достижение компромисса между минимизацией числа элементов " машинных" графем (например, пробелов и точек) и психологической привычностью традиционного шрифта. Во-вторых, шрифты анализируются как специфический объект визуального восприятия, причем здесь существенны не только скорость и надежность распознавания аллографа, но и эмоциональный и эстетический эффект, производимый им на читателя. Продолжением этой работы является традиционная полиграфическая задача размещения текста на странице, что с одной стороны, подвергается осмыслению в лингвистике текста как аспект его гиперсинтаксического членения, а с другой, трансформируется в задачу автоматической структуризации текста (Гринбаум, 1989).

Проблема расположения текста на листе, его членения на абзацы, цветового исполнения приобретают особую актуальность в связи с расширением сервисных возможностей ЭВМ, позволяющих возродить способы представления текста, распространенные до книгопечатания, например, построения сложных схем и таблиц, не только двумерных, но и многомерных (Васюнцов, Котов, 1989). Компьютеризация работ такого рода, в частности использование электронного экрана, порождает сложный комплекс лингвистических, психологических, физиологических и когнитологических проблем (Прохоров, Разлогов, Розин, 1989). Некоторые из них уже обсуждались в прошлом, однако на совершенно иной концептуальной основе, например, в иконоборческих спорах, литургике, культовой архитектуре (т.е. в принципиально синтетических сферах), понимание которых предполагало обращение к сложным герменевтическим процедурам.

Работа с рукописными текстами в рамках прикладной лингвистики связана прежде всего с проблемой его распознавания, а также с решением психологических проблем. Одновременно такие тексты вводят в систему лингвистического знания результаты криминалистических, текстологических и характерологических исследований почерка.* Помимо решения задач, связанных с

* Ср., например, мнение В.Хлебникова о необходимости издания факсимильно воспроизведенных рукописей стихотворений, что отчасти достигнуто многочисленными факсимильными воспро-

идентификацией личности и ее психологического состояния, атрибуции и датировки текста, это направление дает возможность получить некоторые сведения о механизме речевой деятельности, прежде всего при исследовании нарушений письма. Все это позволяет рассматривать рукописный текст как результат сложной деятельности человека, весьма полно и разносторонне выражающей особенности его психологии и физиологии (весьма показательны в этом плане результаты анализа ошибок, сбоев и т.п.).

Особая область - педагогические задачи обучения письму. Это не только путь снабдить человека еще одним средством коммуникации, но и мощное средство воспитания личности в целом. Именно это определяет то внимание, которое уделяется чистописанию при авторитарном или преимущественно эстетическом образовании.

Упомянутые психологические и педагогические аспекты письма органически коррелируют с представлениями о письме как священном действии, в качестве которого оно интерпретируется в герменевтике. К тому же современные данные о передаче навыков при поедании необученной группой червей(планарий)группы обученных показывают, что поедание свитка с текстом при инициации (посвящении) как особом герменевтическом действии не столь уж беспочвенны.*

Следующий вид физической формы речи - устная речь.Будучи представленной звуковыми колебаниями, она, казалось бы, является наиболее естественным, подходящим материалом для создания физикалистских, привычных для науки методов автома-

---

изведениями в последнем собрании его сочинений (Хлебников, 1986, с.619-623).Эта точка зрения Хлебникова по существу относится к поэтике как одному из вариантов деятельностного отношения к языку, коррелирующему с филологией как исследовательской концепцией (Чебанов, Мартыненко, 1990).

*Показательно,что в повседневный рацион питания человека входят нуклеиновые кислоты,являющиеся генетическими текстами и нуклеотиды - буквы этих текстов (Ратнер,1975). С нуклеиновыми кислотами связываются и механизмы долговременной памяти,и,в частности,передача навыков в указанном опыте с планариями.Ср.это с представлением гностика Марка о построении антропоморфного тела божественной истины из букв алфавита.

тической обработки. Однако на протяжении долгого времени в фонетике сосуществовали акустический и фонологический подходы, которые не сопрягались друг с другом. Разработка формантного анализа позволяет заполнить этот пробел. При этом крайне поучительным в методологическом плане является отсутствие прямого соотнесения физических свойств звука и его дифференцирующей роли в качестве языкового знака.*

В свою очередь эти исследования являются базой для широкого круга работ по автоматическому распознаванию и синтезу устной речи. При этом задача распознавания оказывается более сложной, поскольку здесь требуется выделение инварианта в речи разных дикторов, что является прикладным коррелятом упомянутой теоретической проблемы. Другой стороной этой практической задачи является описание речевых особенностей отдельного индивида, диагностика по речи его психического и физиологического состояния. При этом оказывается возможным обсуждать и чисто физиологическое воздействие речи как на говорящего, так и на слушающего, что позволяет понять практику молитв, заклинаний, песнопений (Коломийцева, 1989; Налимов, 1978), концептуально осмысливаемых в герменевтике.

Эстетические аспекты воздействия звучащей речи на человека ныне осмысливаются в лингвистике в связи с проблемой звукового символизма.** На этом пути идет возрождение представлений о том, что по крайней мере иногда связь означаемого и означающего оказывается мотивированной (Воронин, 1984), т.е. возрождается одна из фундаментальных для герменевтики идей о мистике языка. Говоря о мотивированной связи между означаемым и означающим, нужно иметь в виду, по крайней мере два компонента. Первый является собственно семантическим, и

---

* Характер этого соотношения с точки зрения авторов мог бы стать центром кристаллизации лингвистической теории нового типа (см.подробнее Зиндер, 1964). Такие построения в других науках практически отсутствуют. Нечто подобное представляет разве что концепция семейств изоадапторных и изоакцепторных т-РНК (Ратнер, 1975).

** Показательным является появление самого термина "звуковой символизм" - в лингвистике принято говорить не о символах, а о знаках.

именно он рассматривается в современной фоносемантике. Второй связан с ритмом как особой формой организации смысла, что позволяет говорить о семантике ритма (Налимов, 1978). Практическим следствием этого является использование текстов с особой ритмической и акустической организацией в психотерапии, педагогике (ср. стихотворное оформление некоторых правил: "кто и шутя и скоро пожелает ли узнать число, ужъ знаетъ"), ораторском искусстве, вокальной речи (Морозов, 1987).

Другие аспекты исследования устной речи связаны с работами по социолингвистике, криминалистской диагностике, патологии речи и т.п.

Психо- и нейролингвистические исследования устной речи тесно сближают лингвистическую проблематику с задачами исследования организма человека и его мозга. Здесь формируется принципиально важное для понимания онтологии языка представления об инвариантах восприятия, что наиболее полно выражено в представлениях о языках мозга (Прибрам, 1975; Глезер, 1985). Примечательно, что на этом пути ищется связь субстратно представленных физиологических процессов и ментальных по своей природе представлений, т.е. эта проблема вполне аналогична проблеме соотнесения фонетического и фонологического аспектов языка. Не исключено, что этот изоморфизм имеет весьма глубокую природу, поскольку процессы функционирования нервной системы базируются на уже упоминавшихся механизмах использования т-РНК и подобных им механизмах. Таким образом, фонетические исследования, сопрягаемые с нейрофизиологией, начинают смыкаться с изучением плана содержания. Существенно здесь также и то, что исследования по психофизиологии речи заставляют обращаться не только к нервной системе человека, но и к его психосоматической организации в целом и прежде всего к строению и функционированию органов дыхания. Подобные изыскания естественным образом смыкаются с проблематикой пневматологии, актуальной для герменевтики.

Проблематика прикладной лингвистики, связанная с планом выражения, сопрягается с осуществлением Гильбертовской программы построения языка как исчисления, в результате чего механизм языковой рецепции распадается на два: распознавание образа экспонента единицы языка как знакового средства и собственно понимание смысла. При этом оказывается, что при

понимании текста на формальном языке значимы многие не формализуемые проблемы рецепции экспонента языкового знака, в то время как работа со смыслом принципиально оказывается алгоритмизируемой. Там же, где языковое построение многомерно, неформализуемы не только процессы рецепции знака, но и манипулирование со смыслом, что требует привлечения герменевтической практики.

Следующим шагом в развитии прикладной лингвистики было обращение к семантике. Этот шаг был необходим, так как все попытки создать лингвистический автомат, который опирался бы исключительно на план выражения, оказались безуспешными. Обращение к семантике проявилось сначала в попытке зафиксировать смысл значимых компонентов языка. Это вывело на первый план проблему определения понятий, выявления смысловых связей и их структуризации, что, в свою очередь, возбудило интерес к традиционной проблематике классифицирования, составления таблиц, схем, т.е. к тем инструментам, которые появились еще в древности и интенсивно развивались в рамках герменевтических представлений - античных и средневековых. В рамках реальной лингвистической практике использовались лишь простейшие способы представления данных, однако в теоретических изысканиях иногда вспоминали об упомянутых исторических корнях, но только в той мере, в какой это распространялось на концептуальное значение. На практике же оказалось важным учитывать стилистические особенности, коннотации. Более того, только учет коннотативной сферы допускает " подстраивание " лингвистического автомата под текст, а не текста под имеющиеся технические средства (Гончаренко, Шингарева, 1984). На этом пути возникла естественная потребность в обращении к традиционным лингвистическим, филологическим и, конечном счете, герменевтическом представлениям о смысле. Тем не менее, и до настоящего времени объектом лингвистики остаются прежде всего достаточно формализованные, относительно простые тексты, имеющие схематическое содержание - тексты на так называемых ограниченных подъязыках. Всякое расширение класса текстов, доступных для автоматического анализа, требует привлечения лингвистами-прикладниками все более длинного перечня филологических и герменевтических идей. Более того, не будет преувеличением сказать, что оселком, на котором можно испытывать концеп-

туальное и техническое совершенство лингвистического автомата являются поэтические и сакральные тексты. Наиболее общей характеристикой этих текстов является их полисемия. Причем эта полисемия не просто следствие произвола их понимания, а часто результат особой организации самого смысла, заложенного автором. В связи с этим представляется весьма интересным различение четырех смыслов текста Писания в средневековой герменевтике или теория и практика суфийской поэзии (Чебанов, Мартыненко, 1990).

Однако герменевтика отличается от прикладной лингвистики тем, что в последней рассматривается не только иерархическая организация семантики, предполагающая последовательное погружение во все более и более глубокие слои смысла, но и различные способы выявления вариативности смысла текста "горизонтального типа". В последнем случае речь идет о ситуациях, когда один и тот же текст по-разному понимается различными (национальными, профессиональными, территориальными и т.п.) группами читателей, каждая из которых имеет одинаковую (или сопоставимую), но разноаспектную глубину понимания текста.

Отмеченное обстоятельство делает понятным и даже естественным то, что в последние годы на авансцену прикладной лингвистики выходят проблемы прагматики (Новое в лингвистике, 1986). Центральной идеей прагматики является то, что невозможно говорить об одинаковом понимании одного и того же текста разными читателями. Такая позиция во многом противоречит традициям филологии, лингвистики, семиологии, однако она гармонична для герменевтики.

Следует обратить внимание на то, что обращение к прагматике определило интерес лингвистов к референции, значимой в герменевтике. В связи с этим показательно прагматическое осмысление элементов языка, лишенных лексического значения (например, местоименных), которые обеспечивают снятие полисемии текста (Падучева, 1985). Это является операциональным коррелятом герменевтических изысканий о статусе "Я" и Авторе Книги бытия. Принципиальным для рассматриваемой темы оказывается интерес к перформативным высказываниям (Богданов, 1983), что коррелирует с герменевтическим пониманием слова как средства творения. Одним из типов перформативов является вер-

диктив, дающий имя. В связи с этим показательны тотальные переименования, характерные для революционных эпох, и являющиеся способом преобразования человеческого самосознания. Еще одна черта сближения герменевтики и прагмалингвистики – интерпретация молчания как нулевого речевого акта (Богданов , 1986), напоминающая учение о молчании в исихазме.

Другим важным аспектом прагматики является рассмотрение речевой деятельности как сложного вида социо-психо-биологической активности, в которой значимыми оказываются не только физические, химические, биологические, психологические и социальные аспекты, но и то, что речевой акт сам по себе выступает как значимый вид человеческой деятельности (Новое в лингвистике, 1986), когда слово становится действием, делом. Таким образом понимаемая прагматика вбирает в себя практически все результаты, достигнутые психо- и социолингвистикой. Это обстоятельство еще больше сближает прикладную лингвистику с герменевтикой.

Отметим также, что с психологическим аспектом порождения и восприятия речи связана количественная универсалия, известная как гипотеза В.Ингве (Ингве, 1965). Эта гипотеза связывает числовые ограничения на длину регрессивных цепочек с объемом оперативной памяти: число хранимых в ней промежуточных символов не должно превышать миллеровского магического числа 7 ± 2. Аналогичными универсалиями являются числовые ограничения на степень гнездования и ширину синтаксических структур (Хомский, 1966; Шрейдер, 1971; Гладкий, 1985). Эти числовые универсалии перекликаются с арифмологическими представлениями герменевтики.

Итак, можно указать по крайней мере три уровня сближения современной прикладной лингвистики и герменевтики:*

*В соответствии с идеями, изложенными в работе (Чебанов, Мартыненко, 1990), было бы естественным сопоставлять герменевтику с прагмалингвистикой, либо литургику с прикладной лингвистикой. Однако от такого сопоставления по ряду причин можно отказаться. Во-первых, тому, кто не является носителем герменевтической и литургической традиций, различить их крайне трудно. Во-вторых, ввиду того, что для науки в целом характерна известная осторожность при формулировке позитивных утверждений на фоне необходимости опираться даже на предварительные

I. Всеохватность трактовки природы языка и текста, выступающих как результат деятельности целостного человека, живущего в сложно организованном мире. Такая интерпретация требует привлечения идей и методов самых разнообразных научных дисциплин, которые не различены в исходном синкритизме натурофилософского подхода, свойственного герменевтике. Порождение и восприятие текста при этом выступают как акты действия, преобразующие мир или его значительные фрагменты, а речь оказывается в высшей степени полисемичной, допускающей множество трактовок в зависимости от глубины ее постижения или характера деятельностной ориентации читателя. Наряду с этим глобальным сходством существует и фундаментальная противоположность представлений о языке в герменевтике и прикладной лингвистике. В герменевтике всеохватный подход к языку является следствием ориентации мировоззрения на целостность бытия, а в прикладной лингвистике онтология языка не рассматривается вовсе, а его отдельные стороны фигурируют в качестве материала при решении конкретных инженерных задач, требующих междисциплинарного подхода (Богданов, Бондарко и др., 1987). Пользуясь философскими категориями можно сказать, что бытие языка в герменевтике является бытием в себе, а в прикладной лингвистике – бытием для себя. Если предельно заострить эту мысль, то без всякого преувеличения можно сказать, что прикладная лингвистика занимается изучением "трупа" языка,* т.е. изгоняет из языка какое бы то ни было одухотворенное начало. Именно такой подход и обеспечивает возможность машинной переработки текста. Поэтому сходство представлений герменевтики и прикладной лингвистики касается внешних сторон, но за ними стоит совершенно разный смысл. Так, например, упоминавшееся выше внимание к молчанию, свойственное и герменевтике, и прикладной лингвистике, занимает в этих познавательных концепциях совершенно разное место. В исихазме – это такое молчание, которое дает возможность человеку услышать Слово Божие,

---

результаты в практической деятельности, прикладная лингвистика поглощает значительно больший фактический материал, чем прагмалингвистика.

* Ср., у Хлебникова: "жизнь уступила власть союзу трупа и вещи" (Хлебников, 1986, с.191).

т.е. молчание здесь – это отказ от речевого акта, в то время как такой речевой акт в прагмалингвистике является просто разновидностью речевого акта. Однако объектом рассмотрения и в том, и в другом случае является молчание – объект, не свойственный филологии и лингвистике. Вместе с тем, одна из разновидностей молчания как речевого акта сближает герменевтику и прикладную лингвистику. Речь идет о молчании, выражающем уважение к собеседнику. Но главной чертой, сближающей герменевтику и прикладную лингвистику, является структурное богатство рассмотрения языка и речевой деятельности. Прикладная лингвистика – единственная концепция языка, сопоставимая с герменевтикой по тонкости вводимых различений и числу учитываемых опосредованных связей. Такая детальность и глубина структурного анализа определяется тем, что лингвистический автомат лишен сознания и возможности непосредственного восприятия смысла. Именно поэтому в прикладной лингвистике необходим сверхскрупулезный и максимально многоаспектный анализ языка и текста, обеспечивающий алгоритмическую работу с планом содержания на основе анализа плана выражения. В сравнении с таким анализом филологические и лингвистические штудии, ориентированные на человеческое восприятие текста на естественном языке, зачастую кажутся бедными и примитивными.

На фоне этих глобальных сближений и противопоставлений, между герменевтикой и прикладной лингвистикой существуют различия в интерпретации отдельных вопросов, а главное в увязывании между собой разных относительно самостоятельных блоков. При этом для герменевтики характерно в целом более прямое соотнесение разнородных фрагментов (например, дыхания и духа в пневматологии), в то время как в прикладной лингвистике такого рода сближение опосредуется через определенное число промежуточных звеньев (например, через физиологию и психологию речи, психологические аспекты семиотики и т.п.).

2. Сходство отдельных, достаточно крупных концепций прикладной лингвистики и герменевтики. Примерами такого сходства будут: подходы к толкованию полисемии, интерпретация речевой деятельности как действия, типология реципиентов и некоторые другие. Сюда же тяготеет группа проблем, находящихся на грани третьего уровня (см. ниже), такие как проблема звукового символизма, психо-семиотическая значимость графичес-

кого оформления текста и т.п.

3.Совпадение частных, относительно изолированных концепций, таких, как проблема сохранения смысла текста при переводе и введении иноязычных заимствований для достижения этой цели, что делает, в частности, бессодержательным пуризм в отношении языковых заимствований.

Рассмотренные способы интерпретации языка и текста можно суммировать в нижеприведенной таблице, в которой отмечены не только фактическое положение в прикладной лингвистике, но и те порою слабо выраженные тенденции ее развития, которые представляются значимыми в обсуждаемом контексте.

На основании сказанного можно выделить следующие черты, сближающие современную прикладную лингвистику с герменевтикой.

I. Прежде всего налицо размывание границ лингвистики, органическое слияние ее с другими дисциплинами и формирование представлений о языке как объекте междисциплинарных исследований. При этом становится актуальным другой способ расчленения мира на сферы деятельности и в частности выделяется такой слой мира, который связан с языком и речью (предметом какой бы дисциплины он не являлся).

2. Возрождается представление о речи как природном процессе, части мира, в котором произвольность связи означаемого и означающего весьма относительна.

3. Язык соотносится с числом, которое лежит в основе мира (представление арифмологии) или является универсальным средством описания (математические методы прикладной лингвистики).

4. Складывается понимание текста как принципиально многомерного образования (в том числе и в плане выражения, притом что единицы последнего могут быть гетерогенными – буквы и символы, шрифт разного цвета и т.п.).

5.Речевая деятельность начинает рассматриваться как органически связанная с телесно-физиологической организацией человека, соприродная с ним и активно на него влияющая (ср. пневматологию в герменевтике). Это определяет широкое привлечение диагностики физиологического и психического состояния человека по его речи и письму.

6. Выявляется чрезвычайная значимость материала плана

выражения как для функционирования языка, так и для его изучения. Строго говоря, вся эта проблематика относится к внешней лингвистике (по Соссюру) и должна выступать в качестве объекта исследования для особой дисциплины, изучающей речь (лингвистика речи). В связи с этим выявляется особое значение чистописания как средства формирования личности и почерка как способа самовыражения с одной стороны и дидактики и эстетики устной речи с другой.

7. Полисемия начинает выступать как фундаментальная характеристика текста и его единиц, причем очевидным оказывается послойное строение смысла.

8. Коннотации (в широком понимании) выступают как средство послойной интерпретации текста и выявления его символического смысла.

9.Интерпретация текста опирается на сложные классификационные построения,предполагающие многомерные способы представления (таблицы, матрицы, семантические сети и т.д.).

IO.Слово осознается как компонент деятельности, как дело, раскрывая свою созидательную силу в зависимости от типа адресата.

ж ж ж

На основе рассмотренного материала можно сделать следующие выводы:

I. Современное положение науки о языке можно квалифицировать как замыкание цикла исторического развития представлений о языке, когда на очередном этапе развития на авансцену прикладной лингвистики выходят действующие лица, которые более свойственны герменевтике и в определенное время были отброшены как выдумки, фантазии и беспочвенные утверждения.

2. Циклический характер развития представления о языке и речи может выступать как эвристическое средство нащупывания тенденций развития современной лингвистики, что указывает, в частности, на целесообразность уже сейчас обратить серьезное внимание на результаты, достигнутые в герменевтике.

3. Необычайное расширение представлений о природе языка достигается в настоящее время благодаря тому,что лингвистика становится прикладной дисциплиной, а практика требует рассмотрения тех аспектов языка, которые являются дискуссионными для академической традиции.

4.Говоря о сближении прикладной лингвистики и герменевтики, нужно иметь в виду введенные нами три уровня этого сближения. Это необходимо для адекватных экстраполяций, поскольку на фоне сходства общей ориентации и совпадения отдельных деталей в герменевтике наблюдается целостная картина, в то время как в прикладной лингвистике "герменевтические сюжеты" пока представляют собой некоторую свалку.В связи с этим прикладная лингвистика нуждается в формировании общей концепции (она уже начала складываться в рамках прагмалингвистики), и таким образом, именно прикладная лингвистика должна стать источником построения теоретического базиса науки о языке в целом.

5. В настоящее время можно выделить два центра кристаллизации "квазигерменевтических" идей прикладной лингвистики. Во-первых, это осознание значимости плана выражения текста в целом и в частности его материала, который не случаен,а сопричастен психофизической организации человека.Во-вторых,осознание значимости неэксплицированных компонентов семантики,которые приобретают при нынешнем видении языка глобальное значение. Послужив во французской медеевистике основой формирования представления о менталитете, а в методологии науки — представления о парадигме, в области создания искусственного интеллекта это стало основой когнитологии как рефлексии неявного знания, выявление которого путем интервьюирования экспертов является составной частью создания экспертных систем (Экспертные системы,1987).Такая деятельность, являясь по сути экзегетикой,создает общекультурные предпосылки для конституирования герменевтики в качестве своего иного. Примечательно, что оба центра кристаллизации связаны со специфически прикладными задачами – материальной репрезентацией эксплицитно представленного плана выражения и ориентацией на удовлетворение информационных запросов индивидуального пользователя (эволюции теории и практики информационного поиска от требования релевантности к требованию пертинентности).

6. Отсутствие единой концептуальной базы в прикладной лингвистике при богатстве и достаточной полноте привлекаемого материала определяется в значительной мере отсутствием подходящей методологической базы, которая позволяла бы плодотворно рассматривать всю проблематику,касающуюся человека.

Л И Т Е Р А Т У Р А

Богданов В.В. Иллокутивная функция высказывания и перформативный глагол. – В кн.: Содержательные аспекты предложения и текста. – Калинин: Изд-во Калинин. ун-та, 1983, с.27–38.

Богданов В.В. Молчание как нулевой речевой акт. – В кн.:Языковое общение и его единицы. – Калинин: Изд-во Калинин. ун-та, 1986, с.12–18.

Богданов В.В., Бондарко Л.В., Буторов В.Д., Герд А.С. Прикладная лингвистика и теория языка. – В кн.: Структурная и прикладная лингвистика. Межвузовский сборник. Вып.3 – Л.: Изд-во Ленинградского ун-та, 1987, с.3–16.

Воронин С.В. Основы фоносемантики. – Л.: ЛГУ, 1982.

Васюнцов Ю.Д., Котов Ю.В. Основы машинной геометрии и графики. Ч.I.– М.: Московск.текст.ин-т им.А.Н.Косыгина,1989.

Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. – М.: Наука, 1985.

Глезер В.Д. Зрение и мышление. – Л.: Наука, 1985.

Гончаренко В.В., Шингарева Е.А. Фреймы для распознавания смысла текста. Кишинев: Штиица, 1984.

Гринбаум О.Н. Структуризация художественной прозы с использованием ЭВМ (I): формально пунктуационный метод структуризации.–В кн.: Квантитативная лингвистика и автоматический анализ текстов.–Тарту: Уч.зап.Тартуск.ун-та , 1988, с.74–88.

Зиндер Л.Р. Общая фонетика. – Л.: Наука, 1964.

Ингве В. Гипотеза глубины. – В кн.: Новое в лингвистике. Вып.4, М.: Прогресс, 1965, с.141–218.

Коломийцева О.А. Структурно-семантические особенности текстов-заговоров.–В кн.:Семантика и прагматика языковых единиц. – Тюмень, Изд-во Тюменск.гос.ун-та,1989, с.22.

Морозов В.П. Тайны вокальной речи. – Л.: Наука, 1987.

Налимов В.В. Непрерывность против дискретности в языке и мышлении. – Тбилиси: Изд-во Тбилиск.ун-та, 1978.

Новое в лингвистике. Теория речевых актов. Вып. 17. – М.: Прогресс, 1986.

Падучева Е.В. Высказывание и его соотнесенность с действительностью. – М.: Наука, 1985.

Прибрам К. Языки мозга. – М.: Прогресс, 1975.
Прохоров А.В, Разлогов К.Э, Розин В.Д. Культура грядущего тысячелетия. – М.: Вопросы философии, 1989, № 6,с.17–30.
Ратнер В.А. Молекулярно-генетические системы управления. – Новосибирск: Наука, 1975, с.39–115.
Сусов И.П. К предмету прагмалингвистики. – В кн.: Содержательные аспекты предложения и текста. – Калинин: Изд-во Калининск. ун-та, 1983, с.3–15.
Хлебников Велемир. Художники мира. – В кн.: Хлебников Велемир. Творения. М.: Советсткий писатель, 1986, с.619–623.
Хомский Н. Формальные свойства грамматик. – В кн.: Кибернетический сборник.Новая серия. Вып.2 – М.: Мир, 1966,с.121–230.
Хомский Н. Синтаксические структуры. – В кн.: Новое в лингвистике. Вып. 2 – М.: Иностранная литература, 1962,с.412 – 527.
Чебанов С.В., Мартыненко Г.Я. Основные типы представлений о природе языка. В кн.: Linguistica. Исследования по общему и прикладному языкознанию. Тарту: Тартуск.гос.ун-т, 1990.
Шаров А.А. Зимняя школа по биосемиотике – Журнал общей биологии, 1990, № 2, с.291–293.
Шрейдер Ю.А. Равенство. Сходство. Порядок. – М.: Наука, 1971.
Экспертные системы. Принципы работы и примеры. Под ред.Р.Форсайта. – М.:"Радио и связь", 1987.

Аспекты сближения герменевтики и прикладной лингвистики

	Герменевтика	Прагмалингвистика	Аспект сближения герменевтики и прагмалингвистики*
I	2	3	4
	О Н Т О Л О Г И Ч Е С К И Й С Т А Т У С		
Тип семиологических средств	символ	комплекс знаковых средств	принятие возможности существования символов
Размерность отношения к языку	бесконечномерность	"нужномерность"	многомерность
Фундаментальное время	панхроническое	физическое	универсальность времени
Характер понимания	бесконечное углубление	понимание разных слоев смысла	многослойность понимания
Цель обращения к тексту	спасение	осуществление действия	деятельностная ориентация
Отношение к числу	арифмология	квантитативные методы	сближение онтологии числа и слова
Статус текста	часть мира	компонент деятельности	сходство онтологического статуса
Отношение между человеком и языком	дар речи (двусторонний)	средство практической деятельности	язык - основное средство деятельности человека в мире
Представление целостности	духовная целостность	целостность актов речевой деятельности	конструирование целого из разнородных компонентов
Семиотическая ориентация	прагматика	прагматика	ориентация на прагматику
Универсальная семиотическая модель	текст бытия	семантическая сеть	многомерность представления семантики
	С Е М И О Л О Г И Ч Е С К И Й С Т А Т У С		
Роль референта	сакрален	интересен	значимость референта

*Имеется в виду деятельностный коррелят прагмалингвистики - прикладная лингвистика (см.примечание на с.9).

I	2	3	4
Связь означаемого и означающего	сакрализованная природа	гетерогенная детерминация	возможность существования мотивированных связей
Статус устной речи	проявление духа	определяется природой человека	связь с психофизиологией дыхания, ориентация на время
Статус письменной речи	явление тайного (в том числе профанация)	визуализированная форма существования речи	связь с психологией зрения, ориентация на пространство
Статус материала плана выражения	священен	неслучаен	неслучайность выбора
Роль этимологического значения	сверхэффективность (перформативность)	средство различения синонимии	возрождение роли этимологического и псевдо - этимологического значения
Роль коннотаций	неразделимость концептуального значения и коннотаций	имплицитное средство актуализации иллокутивных интенций	значимость коннотаций
Роль омонимии	проникновение в замысел	поиск средств разрешения омонимии на глубинном уровне	проникновение в смысл текста
Контекстная зависимость смысла и знаковых средств	значимое приращение смысла в контексте	взаимодействие контекстного и инвариантного значения	значимость контекста
ДЕЯТЕЛЬНОСТНЫЙ СТАТУС			
Деятельностный коррелят научной концепции языка	литургика	прикладная лингвистика	стремление донести смысл до адресата
"Присущесть" человеку	бытийная	психофизиологическая	глубокое родство человека и языка

1	2	3	4
Соотношение "слова" и "дела"	средство творения	слово есть дело	слово как орудие деятельности
Тип адресата	"Имеющий уши"	разные типы адресатов	ориентация на "имеющих уши"
Роль перформативов	средство творения	орудие деятельности	средство преобразования внеязыковой реальности
М Е Т О Д О Л О Г И Ч Е С К И Й С Т А Т У С			
Методологический подход	натурофилософский подход (онтологизм)	системно-деятельностный подход, комплексный подход	методология соединения разнородных объектов
Базовый метод	откровение	экспертиза	использование ненормативных средств
Объект	текст как вместилище смысла	процесс восприятия и порождения текста	задание объекта через план выражения
Предмет	сакральный смысл	речевое общение	процессуально-событийная организация текста

IDEAS OF HERMENEUTICS IN APPLIED LINGUISTICS

S.V. Chebanov, G.Ya. Martynenko

S u m m a r y

The present situation in the science of language is qualified by the authors as a closing of a cycle of historical development of cognitive conceptions, with the linguistics' proscenium being taken by the characters that are more peculiar to hermeneutics. However, against a background of similarity of a general orientation and coincidence of separate details, in hermeneutics there is observed a more integral picture, while in applied linguistics hermeneutic "plots" are still representing some kind of a dump. That's why applied linguistics is standing in need of development of a general conception which elements started forming in pragmatic linguistics. In conclusion it is noted that applied linguistics is turning into a source of creating a theoretical basis of the science of language as a whole.

Х Р О Н И К А

ВСЕСОЮЗНАЯ КОНФЕРЕНЦИЯ ПО КОМПЬЮТЕРНОЙ ЛИНГВИСТИКЕ

С 29 по 31 мая 1990 года в Тартуском университете проводилась I-я Всесоюзная конференция "АКТУАЛЬНЫЕ ПРОБЛЕМЫ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ", организованная Тартуским университетом совместно с Московским государственным университетом, Институтом языкознания АН СССР, Институтом языка и литературы АН Латвии. Значительный вклад в подготовку и проведение конференции внесла Межвузовская проблемная группа "Текст как объект междисциплинарных исследований".

В конференции принимали участие 93 ученых из 36 разных городов Советского Союза, из них 15 докторов наук и 51 кандидат наук. На 2 пленарных и 6 секционных заседаниях было заслушано 70 докладов. В конце конференции было проведено заседание за круглым столом на тему "Человеко-машинное общение: качественные и количественные аспекты".

На конференции был обсужден широкий круг актуальных проблем компьютерной лингвистики, прежде всего в связи с интеллектуализацией вычислительных систем, инженерно-лингвистическим моделированием, машинным переводом и автоматизацией лексикографических исследований, а также созданием и совершенствованием экспериментальных и промышленных систем анализа и синтеза текста.

На первом пленарном заседании, посвященном теме "Теоретико-методологические аспекты компьютерной лингвистики" и проведенном в актовом зале Тартуского университета, выступил со вступительным словом и приветствием от имени руководства университета профессор ТУ М.А. Шелякин. Затем были заслушаны доклады об общих проблемах и перспективах компьютерной лингвистики (А.А. Поликарпов, Ю.А. Тулдава) и о других актуальных теоретических и практических аспектах КЛ (Л.Л. Нелюбин, Э.И. Королев, Р.Ю. Кобрин).

Тематика секционных заседаний охватывала следующие общие проблемы: лингвистические текстовые процессоры (с докладами выступили Н.П. Дарчук, Т.А. Грязнухина, С.А. Яблонский, Ж.Г. Мошкович и др.); автоматический анализ учебных, поэтических и музыкальных текстов (Я.А. Микк, Е.М. Брейдо, Е.А. Цыб, В.К. Детловс и др.); автоматизированные лексикографические системы и анализ компьютерных словарей (А. А. Поликарпов, Л.И. Колодяжная, С.В. Лесников, Ф.А. Циткина,

В. Жилинскене и др.); машинное моделирование естественно-языкового общения: лингвистические и когнитивные аспекты (Х.Я. Нйм, М.Э. Койт, М.М. Ференц-Мошинская, С.Г. Фунштейн и др.); различные проблемы автоматической классификации и автоматического моделирования в лингвистике (Г.Г. Сильницкий, Ю.А. Тулдава, И.В. Бахмутова, А.Х. Джубанов, Э.В. Калашников, В.Е. Остапенко, П.В. Сороколетов и др.); машинный перевод и машинные фонды языков (В.А. Дризуле, Н.Ф. Клименко, Е.А. Карпиловская, О.Г. Капанадзе и др.).

На заключительном пленарном заседании на тему "Теоретические модели языка и их реализация на ЭВМ" выступили В.И. Перебейнос ("Стилеразличительная способность зон связей"), Ю.К. Крылов и А.А. Поликарпов ("Реализация на ПЭВМ системного анализа комплекса статистических характеристик лексики") и Ю.К. Крылов ("Волновая теория, временные ряды и проблемы статистических оценок целостности художественного произведения").

После заключительного пленарного заседания была проведена общая дискуссия и подведены итоги конференции.

В итоговых документах конференции отображены следующие решения и рекомендации, принятые участниками конференции.

Примерами разработанных и представленных на конференции систем КЛ отметились: система автоматического морфологического и синтаксического анализа текста (Институт языковедения АН УССР). Словарный процессор (Институт русского языка АН СССР), Процессор русского языка АСПЕРА (ВЦП ГКНТ и АН СССР). Процессор "РУССИКОН-I" (ЛИИЖТ), Система статистического анализа лексической информации (Лаборатория компьютерной лингвистики НПА), Система человеко-машинного естественно-языкового интерфейса (Тартуский ун-т).

Теоретическая проблематика была сосредоточена на актуальных вопросах построения теоретических моделей языка и их реализации на ЭВМ, исследованиях статистических закономерностей организации и функционирования языка, принципиальных теоретико-методологических аспектах компьютерной лингвистики, вопросах построения систем искусственного интеллекта, включая лингвистические и когнитивные аспекты человеческого общения (Московский, Тартуский, Горьковский университеты). На основе обсуждения этих вопросов можно сделать вывод о сближении и о взаимовлиянии работ по искусственному интеллекту и по компьютерной лингвистике.

Конференция констатирует, что лингвистические процессоры постоянно прогрессируют в сторону специализированных интеллектуальных систем. С другой стороны, снабжаясь все более обширным лингвистическим обеспечением, проблемно-ориентированные диалоговые системы становятся все более и более универсальными. В этом же направлении прогрессирует и устройство блока естественно-языкового интерфейса автоматизированных ИПС.

Распространение персональной вычислительной техники и разработка методологии и технологии использования ПЭВМ в лингвистических исследованиях и разработках позволяет автоматизировать лексикографические и другие лингвистические исследования, включая работы по созданию Машинных фондов русского языка и других языков (Институт русского языка АН СССР, Институты языкознания Украины, Эстонии, Латвии и др.). Работы этого плана, кроме того, что они имеют важное прикладное значение, обладают и фундаментально-научной и культурной значимостью, способствуют прояснению принципиальных моментов устройства языка и общения, способствуют сохранению национальных культур.

Вместе с тем конференция констатирует, что данные исследования и разработки, имеющие важное научное и культурное значение, не обеспечены в достаточной степени вычислительной техникой и средствами скоростного ввода и вывода данных, слабо координированы в масштабах страны, что сдерживает их дальнейшее развитие.

Конференция приняла следующие решения:

I. Отметить высокий теоретический уровень разработок по текстовым и словарным лингвистическим процессорам в Институте языкознания им. А.А. Потебни АН УССР, МГУ им. М.В. Ломоносова (Лаборатория прикладной лингвистики), Научно-производственной Ассоциации (НПА) г. Москвы (Лаборатория компьютерной лингвистики), Тартуском университете (Группа прикладной и компьютерной лингвистики), ЛГПИ им. А.И. Герцена, ЛЭТИ им. Ульянова-Ленина, ЛИИЖТ (Лен. ин-т инж. ж/д транспорта) и др.

2. Головной организации страны по разработке лингвистических процессоров - Всесоюзному центру переводов ГКНТ СССР и АН СССР - учесть отмеченный уровень разработок по текстовым и словарным лингвистическим процессорам при координации работ по этой тематике в рамках Государственной программы информатизации общества ИНФОРМАТИЗАЦИЯ-2005, принятой ГКВТИ

СССР (Госкомитет по вычисл. технике и информатике).

3. Обратить внимание ГКВТИ СССР, ГКНТ СССР и ГОСКОМОБР СССР на необходимость государственного приоритетного финансирования указанных организаций и обеспечения их современной вычислительной техникой, прежде всего профессиональными ПЭВМ типа IBM PC/AT и аппаратурой скоростного ввода и вывода данных. Недостаток технических средств сдерживает фундаментальные и прикладные исследования в указанной области.

4. С целью улучшения координации научно-исследовательских работ рекомендовать открыть при МГУ им. М.В. Ломоносова исследовательский и информационно-консультативный центр по компьютерной лингвистике, одной из важнейших задач которого был бы сбор и распространение информации о ведущихся в Советском Союзе и за рубежом работах по компьютерной лингвистике.

5. Считать целесообразным проведение в расширенном масштабе работ по созданию машинных версий словарей русского и других языков (словообразовательных, словосочетаний, синонимов, единиц измерений и пр.), по оптимизации обучения русскому и другим языкам, созданию лингвистической базы в виде машинных фондов для теоретических исследований и разработки действующих систем обработки текстовой и словарной информации.

6. Считать особо перспективной проблему сопряжения текстовых и словарных процессоров с большими текстовыми и словарными базами данных в комплексных автоматизированных системах обработки лингвистических данных.

7. Считать необходимым начать периодическое издание по компьютерной лингвистике. Базой для издания предложить Московский государственный университет.

8. Одобрить инициативу группы московских и ленинградских ученых по изданию всесоюзного журнала по теории и практике современного языкознания (под ред. проф. Л.Л. Нелюбина).

9. В связи со значительными результатами, достигнутыми на конференции, выходом ученых Тартуского университета на всесоюзную тематику, ходатайствовать перед ГОСКОМОБР СССР и Министерством просвещения Эстонии об усилении компьютерной базы Группы прикладной и компьютерной лингвистики Тартуского университета (председатель проф. Ю.А. Тулдава) для обеспечения оптимального использования творческого потенциала

ученых университета и создания лингвистической компьютерной лаборатории.

IO. Отметить существенный вклад, внесенный проблемной группой "Текст как объект междисциплинарных исследований" в развитие исследований в области прикладной и компьютерной лингвистики в Советском Союзе (семинары, конференции и выпуск периодического издания "Квантитативная лингвистика и автоматический анализ текстов" в период 1985-1990 гг.).

II. Основные выводы конференции опубликовать во всесоюзных журналах "Научно-техническая информация, Серия 2" и "Филологические науки".

I2. Отметить высокий уровень организации Тартуским университетом I-й Всесоюзной конференции "Актуальные проблемы компьютерной лингвистики".

I3. Рекомендовать организовать следующую, 2-ю Всесоюзную конференцию по компьютерной лингвистике в 1992 году в Москве на базе МГУ.

А.А. Поликарпов, Ю.А. Тулдава

R E V I E W

AN ALL-UNION CONFERENCE ON COMPUTATIONAL LINGUISTICS

A. Polikarpov and J. Tuldava

S u m m a r y

An All-Union Conference "Actual Problems of Computational Linguistics" was held in Tartu on May 29-31, 1990. Two plenary meetings were devoted to theoretical-methodological aspects of computational linguistics. Six sections dealt with (1) linguistic text processors; (2) automatic analysis of educational, poetic, and musical texts; (3) automatized lexicographic systems; (4) machine-aided modelling of natural-language communication: linguistic and cognitive aspects; (5) automatic classification in linguistics; (6) machine translation and computerized language funds. A Round Table discussion was also held on qualitative and quantitative aspects of man-machine communication.

Р Е Ц Е Н З И Я

О ЗАКОНЕ МЕНЦЕРАТА

Рецензия на книгу: Gabriel Altmann, Michael H. Schwibbe. Das Menzerathsche Gesetz in informationsverarbeitenden Systemen/ Mit Beiträgen von Werner Kaumanns, Reinhard Köhler und Joachim Wilde. - Hildesheim; Zürich; New York: Georg Olms Verlag, 1989. - 132 S.

В рецензируемой коллективной монографии обсуждается действие в разных областях науки т.н. закона Менцерата, который гласит: чем больше целое ("конструкт"), тем меньше его составные части ("компоненты"). Этот закон, или принцип, был впервые - в более узком значении и в виде гипотезы - сформулирован немецким ученым П. Менцератом[+], который при исследовании фонетической структуры немецкого слова обнаружил, что с увеличением длины слов их составляющие части - слоги - в среднем уменьшались в длине. Более широкую лингвистическую трактовку и математическую экспликацию этого явления дал впоследствии известный специалист по квантитативной лингвистике Г. Альтманн, профессор Бохумского университета (ФРГ)[++]. В рецензируемой монографии закон Менцерата получает еще более широкую трактовку. В работе показано, что, кроме лингвистики, действие закона прослеживается также в психологии, социологии приматов и молекулярной биологии, т.е. в областях, где в той или другой форме имеют место информационные процессы. На основе многочисленных примеров из названных областей науки авторам монографии удается демонстрировать универсальный характер закона Менцерата, т.е. принципа обратной пропорциональности целого и его составных частей, причем в качестве составных частей могут выступать не только материальные части целого, но и разные функции или значения, а измерением служит "сложность" в самом общем смысле.

В первых главах монографии, написанных Г. Альтманном, раскриваются основные понятия предлагаемой концепции, ведущей к утверждению упомянутого принципа обратной пропорциональности в качестве специфической закономерности в функцио-

---

[+] Menzerath, P. Die Architektonik des deutschen Wortschatzes. - Bonn: Dümmler, 1954.

[++] Altmann. G. Prolegomena to Menzerath's Law // Glottometrika 2. - Bochum: Brockmeyer, 1980. - Pp. 1-10.

нировании определенного рода социальных систем. Проводится аналогия с известным законом Ципфа-Мандельброта, который также действует не только в лингвистике, но и в других системах, где происходят процессы обмена и хранения информации. Г. Альтманн высказывает предположение, что оба закона (Ципфа-Мандельброта и Менцерата) могут быть следствием некоторого общего, их объединяющего принципа, например, принципа "наименьшего усилия", или какого-нибудь другого глобального принципа, который еще ждет своего раскрытия.

Остальные главы посвящены формулировке лингвистических и других гипотез в качестве следствий из закона Менцерата и проверке состоятельности этих гипотез с помощью статистических методов. Сосредоточимся на некоторых моментах, особенно важных с точки зрения квантитативной лингвистики и общей методологии вывода и интерпретации квантитативных закономерностей в разных областях науки.

При выводе закона Менцерата авторы коллективной монографии представляют три разных, но дополняющих друг друга подхода.

Г. Альтманн, основываясь на предполагаемой обратной пропорциональности величин конструкта (x) и компонента (y), например, длин слова и составляющих его слогов, выражает это отношение в виде дифференциального уравнения, в котором относительный темп изменения величины y, т.е. первая производная y по отношению к самой величине y, символически y'/y, ставится в обратную пропорцию к величине x:

$$\frac{y'}{y} = \frac{b}{x} .$$

Интегрируя, получаем $\ln y = b \ln x + c$ и затем

$$y = ax^b, \qquad (I)$$

где $a = e^c$ и $b < 0$. Это - степенная функция, отражающая монотонное уменьшение y при постоянном увеличении x (на графике - гиперболическую кривую).

В дальнейшем следует принять во внимание дополнительные ("ограничивающие") факторы, которые в реальных ситуациях могут повлиять на связь между x и y (например, позиция компонента в структуре конструкта). Г. Альтманн предлагает в качестве одного из вариантов следующую модификацию исходного дифференциального уравнения:

$$\frac{y'}{y} = \frac{b}{x} + c ,$$

из которого получается обобщенная функция

$$y = ax^b e^{cx} . \qquad (2)$$

Так как при выводе функции (2) дополнительный фактор явно не указан (им может быть b или c), то при b = 0 получается экспоненциальная функция

$$y = ae^{cx} . \qquad (3)$$

Если же в функции (2) принять c = 0, то получается первый вариант – степенная функция (I).

Другие возможности вывода и интерпретации закона Менцерата представляют М. Х. Швиббе и Р. Кёлер. Первый из них связывает соотношение величин конструкта и его компонентов с некоторыми особенностями памяти и рассматривает закон ("правило") Менцерата в качестве модели психической переработки информации. Р. Кёлер, автор известной книги о "языковой синергетике"+, рассматривает закон Менцерата в связи с распределением "структурной информации" в языковом процессе, и при аналитическом описании связи между необходимым объемом памяти ("регистра" для обработки структурной информации компонентов) и числом компонентов он приходит к дифференциальным уравнениям и функциям, аналогичным уравнениям и функциям, предложенным Г. Альтманном.

Основная часть монографии посвящена разбору практических примеров с целью проверки состоятельности закона Менцерата в той или другой конкретной форме, в частности в форме функции (I) при рассмотрении большого числа различных явлений в лингвистике и других областях науки.

Например, на основе данных из разных языков анализируются (Г. Альтманном) следующие связи: длина предложения (как конструкт) – длина фраз (частей предложения, количество которых определяется в данной работе количеством финитных глаголов в предложении); длина фразы – длина слов; длина слова – (фонетически измеряемая) протяженность слогов или звуков в слове; длина слова – среднее число значений. Во всех случаях зависимость длины компонента от длины конструкта хорошо аппроксимируется функцией (I) на уровне статистической значимости не выше 0,05 (статистическая достоверность проверяется с помощью F-критерия Фишера). Применяемая функция удобна не только своей простотой, но и тем, что ее параметрам в рамках закона Менцерата удается дать содержательную

+ Köhler, R. Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. – Bochum: Brockmeyer, 1986.

интерпретацию. Параметр а выражает среднюю длину компонента при начальной длине конструкта (например, среднюю длину слога в односложных словах). Параметр в (при в < 0) указывает на темп уменьшения длины компонента в связи с увеличением длины конструкта. Оба параметра могут служить типологически дифференцирующими показателями при сопоставлении данных из разных языков. Покажем это на одном примере, приведенном в главе о зависимости длины слога от длины слова (с. 51 и след.). Сопоставим значения параметров функции (I) в двух языках:

английский а = 4,09; в = 0,36;
итальянский а = 2,66; в = 0,11.

Значение а указывает на бóльшую среднюю длину односложного слова в английском языке по сравнению с итальянским, а большее значение в в англи́йском языке говорит о более резком уменьшении длины слога в переходе от односложных слов к многосложным (например, в английском языке длина слога в односложном слове в среднем 4,09 фонемы, в двусложном слове – 3,19 фонемы; в итальянском языке 2,66 и 2,46, соответственно).

Интересен раздел, в котором обсуждаются неявные следствия закона Менцерата в эволюционных процессах языка. Г. Альтманн показывает, как тенденция к уменьшению длины слога во многосложном слове может провести к редукции согласных, в частности к частичной редупликации (типа $C_IГ/C_IГC_2$ вместо полной редупликации $C_IГC_2/C_IГC_2$, где С – согласный, Г – гласный звук), а также к ассимиляции, контаминации и т.п. (хотя при таких явлениях могут играть роль и частота употребления, значение слова и некоторые другие факторы). В разделе рассматриваются и другие возможные изме́нения в языке, которые можно приписать действию закона Менцерата.

Далее, в монографии приводится оригинальный эмпирический материал по молекулярной биологии (авторы: Й. Вильде и М. Х. Швиббе), в частности по сравнению величины DNA-носителя генетической информации – с величиной хромосом (компонентов DNA), где наблюдается обратная пропорциональность величин конструкта и компонентов, а также материал по социальным структурам у приматов (авторы: В. Кауманнс и М. Х. Швиббе) при сравнении величины популяции с величиной групп, составляющих популяцию, при сравнении величины группы с активностью и вариацией внутригрупповых интеракций и др. И здесь действует закон Менцерата в своей исходной форме и может быть выражен функцией (I).

Мы не ставили задачу рассмотреть все наблюдения и материалы, содержащиеся в рецензируемой коллективной монографии, а также некоторые дискуссионные методологические проблемы, обсуждаемые в книге самими авторами (формы параметризации, направление детерминации "конструкт – компоненты" и др.). Но и сказанное позволяет заключить, что перед нами глубокое исследование важной закономерности, связанной с эволюцией и функционированием определенного класса сложных систем. Действенность обсуждаемого "закона Менцерата" в лингвистике – в фонетике, морфологии, синтаксисе –, а также в некоторых областях биологии, психологии и социологии, в рецензируемой книге убедительно доказана. Не подлежит сомнению, что коллективная монография свидетельствует о глубокой компетенции авторов в исследуемой теме и умении пользоваться выбранными методами, делать обоснованные выводы. Заслуживает положительной оценки и стремление авторов везде давать содержательную интерпретацию результатов квантитативного анализа. Практическую пользу получает читатель от раздела, где обсуждаются и объясняются статистические методы, применяемые в книге; практически полезно также приложение, в котором приводятся готовые прикладные программы для вычислений на ПЭВМ (по определению параметров функций, по дисперсионному анализу и др.).

Насыщенные фактами и интересными выводами коллективное исследование представляет несомненный интерес для широкого круга лингвистов и специалистов других областей науки. Специфика глубоко проанализированных лингвистических данных из разных языков в ансамбле всей книги соотнесена с закономерностями более общего порядка с целью построения в дальнейшем теории языка в рамках общей теории самоорганизующихся систем. Монографии как целому это придает особую масштабность.

Ю. Тулдава

Review of G. Altmann, M.H. Schwibbe. Das Menzerathsche Gesetz in informationsverarbeitenden Systemen / Mit Beiträgen von W. Kaumanns, R. Köhler und J. Wilde. - Hildesheim; Zürich; New York: Georg Olms Verlag, 1989 (Menzerath's Law in Systems of Information Processing)

Juhan Tuldava

S u m m a r y

The monograph under review deals with the effect of the so-called Menzerath's Law which states: The greater the whole ("construct") the smaller the parts ("components"). It has been shown that this law can be expressed analytically by the power function $y = ax^b$ where y is the length of the component and x - the length of the construct, a and b are constants ($b < 0$). This has been demonstrated on rich material from linguistics - phonology, morphology, syntax (e.g. connection between the length of word and the length of its syllables), and from some fields of biology and sociology.

С О Д Е Р Ж А Н И Е

Андреевская А.В. Квантитативное исследование полисемии корневых слов русского языка XI-XX веков. 3-II

Блехман М.С. Методы автоматической атрибуции документов: практические результаты..... I2-20

Голубева-Монаткина Н.И. Статистические характеристики коммуникативных свойств вопросов и ответов русской диалогической речи..... 2I-3I

Гороть Е.И. Изоморфные и отличительные черты морфемы и слога в распределении длины..... 32-36

Зубов А.В. Системы автоматизации научных исследований в филологии..... 37-54

Иванюк В.Ю., Левицкий В.В. Избирательность сочетания смыслов и возможные способы ее статистического выражения..... 55-6I

Манасян Н.С. Еще раз о дифференциации типов английского научно-технического текста..... 62-66

Остапенко В.Е. Принципы формального решения проблемы соотношения между термином и словом..... 67-77

Савчук С.О. О некоторых содержательных характеристиках стиля..... 78-9I

Чебанов С.В., Мартыненко Г.Я. Идеи герменевтики в прикладной лингвистике..... 92-III

Х р о н и к а :

Поликарпов А.А., Тулдава Ю.А. Всесоюзная конференция по компьютерной лингвистике..... II2-II6

Р е ц е н з и я :

Тулдава Ю. Рец. на кн.: G. Altmann, M.H. Schwibbe. Das Menzerathsche Gesetz in informationsverarbeitenden Systemen. Hildesheim; Zürich; New York: Georg Olms Verlag, 1989 (Г. Альтманн, М.Х Швиббе. Закон Менцерата в информационных системах). II7-I22

SUMMARIES - RÉSUMÉS

Andreewskaya A.W. Russian XI-XX Centuries Root-words' Quantitative Analysis..... 11

Blekhman M.S. Some Methods of Automatic Text Attribution: Practical Results..... 20

Golubeva-Monatkina N.I. Statistical Characteristics of Communication Properties of Questions and Answers of Russian Dialogic Speech..... 31

Gorot' E.I. Isomorphous and Distinguishing Features of Morphemes and Syllables in Their Distribution according to Their Length..... 36

Zubov A.V. A System of Automatic Scientific Research in Philology..... 54

Ivanyuk V.Yu., Levitsky V.V. The Selectivity of Sense Collocation and Possible Ways of Its Statistical Expression..... 61

Manasyan N. Once Again on the Differentiation of English Technological Texts..... 66

Ostapenko V.E. Principes de solution formelle du problème de correlation entre un terme et un mot..... 77

Savchuk S.O. On Some Substantial Characteristics of Style..... 91

Chebanov S.V., Martynenko G.Ya. Ideas of Hermeneutics in Applied Linguistics..... 111

S u r v e y :

Polikarpov A., Tuldava J. All-Union Conference on Computational Linguistics in Tartu (May 29-31, 1990) 116

R e v i e w :

Tuldava J. Review of: G. Altmann, M.H. Schwibbe. Das Menzerathsche Gesetz in informationsverarbeitenden Systemen / Mit Beiträgen von W. Kaumanns, R. Köhler und J. Wilde. - Hildesheim; Zürich; New York: Georg Olms Verlag, 1989..... 122