

4и (Нем)

К584

✓

МИНИСТЕРСТВО
ВЫСШЕГО И СРЕДНЕГО СПЕЦИАЛЬНОГО ОБРАЗОВАНИЯ БССР
МИНСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ ИНСТИТУТ
ИНОСТРАННЫХ ЯЗЫКОВ

На правах рукописи

КОЖЕНЕЦ Тамара Федоровна

ЭНТРОПИЯ НЕМЕЦКОГО ПЕЧАТНОГО ТЕКСТА

10.02.04. — германские языки
(Диссертация написана на русском языке)

А В Т О Р Е Ф Е Р А Т

диссертации на соискание ученой степени
кандидата филологических наук

Минск 1978

МИНИСТЕРСТВО ВЫСШЕГО И СРЕДНЕГО СПЕЦИАЛЬНОГО ОБРАЗОВАНИЯ БССР
МИНСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ ИНСТИТУТ
ИНОСТРАННЫХ ЯЗЫКОВ

На правах рукописи

КОЖЕНЕЦ Тамара Федоровна

ЭНТРОПИЯ НЕМЕЦКОГО ПЕЧАТНОГО ТЕКСТА

10.02.04 - германские языки

/ Диссертация написана на русском языке/

А В Т О Р Е Ф Е Р А Т

диссертации на соискание ученой степени
кандидата филологических наук

Минск-1973

1693

Библиотека
МГПИИЯ

Работа выполнена на кафедре лексикологии немецкого языка Минского государственного педагогического института иностранных языков.

Научный руководитель - доктор филологических наук, профессор Р.Г. ПИОТРОВСКИЙ.

Научный консультант - кандидат филологических наук, доцент Г.П. БОГУСЛАВСКАЯ

Официальные оппоненты:

доктор филологических наук В.В. МАРТЫНОВ

кандидат филологических наук А.Н. ШАРАНДА

Ведущее учебное заведение - Институт языкознания АН УССР, сектор структурной, математической и прикладной лингвистики.

Автореферат разослан "6" марта 1973 г.

Защита диссертации состоится "10" апреля 1973 г. на заседании Ученого Совета Минского государственного педагогического института иностранных языков, г. Минск, ул. Захарова, 21.

С диссертацией можно ознакомиться в библиотеке института.

Отзывы направлять по адресу: г. Минск, 22662, ул. Захарова, 21, Ученому секретарю.

УЧЕНЫЙ СЕКРЕТАРЬ СОВЕТА

Реферируемая работа посвящена количественному исследованию синтаксической / селективной/ и смысловой информации немецкого печатного текста как на уровне фигур, так и на уровне значимых единиц.

Предполагается получить информационно-статистические характеристики немецкого слова, немецкого печатного текста в целом, а также определить количественную оценку смысловой информации некоторых элементов текста через измерения синтаксической информации. При этом предполагается рассмотреть следующие частные вопросы:

1. Определение энтропии и избыточности немецкого печатного текста относительно языка в целом и его четырех основных стилистических разновидностей: публицистической, беллетристической, научно-технической и устно-разговорной в ее письменной фиксации. Эти данные должны быть сопоставлены с аналогичными оценками других языков, полученными тем же путем.

2. Определение зависимости между величиной энтропии и долей: а/ достоверных продолжений; б/ согласных; в/ гласных и г/ пробелов относительно общего числа букв, приходящихся на данную буквенную позицию.

3. Построение информационной лексической схемы немецкого текста.

4. Определение информационной структуры немецкого слова в его буквенном, слоговом и морфемном выражении.

5. Построение информационной схемы немецкого сложного слова на уровне составляющих его основ.

6. Оценка количественного и качественного воздействия лексико-грамматического контекста на информационную структуру немецкого слова.

7. Определение количественной оценки грамматической / флексивной / информации относительно общей лексико-грамматической информации в тексте.

8. Определение количественной оценки селективной и смысловой информации предшествующего контекста.

Решение данных задач осуществляется вероятностно-статистическим методом Шеннона-Пиотровского.

Исходным моментом данного метода является представление языка в виде кода с вероятностными ограничениями: речь / текст /, состоящий из последовательности дискретных единиц / букв, слогов, морфем, основ и т.д. /, выступает в качестве сообщения, где реализуется этот код. Появление каждой такой дискретной единицы в тексте трактуется как последовательное осуществление отдельных опытов. Правильный выбор лингвистической единицы снимает неопределенность, которая существовала до его реализации. Количество неопределенности в системе выборов лингвистической единицы, или иначе энтропия, численно равна величине информации, появляющейся в результате снятия этой неопределенности. Для статистического определения понятия " количество информации " мы используем вероятностный подход¹, т.е. лингвистические

I/.См.:А.А.Марков. Об одном применении статистического метода. Изв.ИАН, 1916, сер.6; К.Шеннон. Работы по теории информации и кибернетике. Перевод с англ. М., 1963

элементы, с помощью которых передается информация, рассматриваются как случайные переменные, обладающие различной вероятностью. Предполагается, что в языке-коде заданы ограничения как на сочетаемость лингвистических единиц речи, так и на их вероятность появления.

Основным моментом метода Шеннона-Пиотровского является проведение психолингвистического эксперимента по определенной процедуре^I с последующей математической обработкой данных эксперимента, оценкой достоверности полученных результатов и их лингвистической интерпретацией.

Первая глава

Основная задача первой главы состоит в определении энтропии и избыточности немецкого печатного текста относительно языка в целом и его четырех функционально-стилистических разновидностей: беллетристической, публицистической, научно-технической и устно-разговорной, а также сравнение полученных результатов с аналогичными данными других языков.

Материалом для данного исследования послужили отрывки 200 оригинальных текстов указанных стилистических разновидностей современного литературного немецкого языка, причем

I/. О подробном описании проведения эксперимента см.:
Р.Г. Пиотровский. Информационные измерения языка.
Л., 1968, стр. 3-47

выборка по каждому стилю представлена 50 текстами длиной в 200 букв каждый.¹

В эксперименте участвовало 26 информантов-носителей языка, все с высшим образованием, двое из них - специалисты в области немецкой литературы и имеют ученую степень.

Основные итоговые данные об энтропии и избыточности четырех стилистических разновидностей общезыковой выборки представлены в таблице I.

Сопоставление величин энтропии и избыточности по стилям немецкого языка показало, что наибольшей избыточностью R /соответственно наименьшей энтропией H /обладает научно-технический стиль. Здесь язык характеризуется "чрезвычайной отработанностью не только терминологии, но и всей манеры изложения, где употреблен ряд постоянно повторяющихся структур, даже штампов, как бы заранее навязывающихся информанту".²

Основные стилистические черты научно-технической речи, как "логичность, объективность, абстрактность, точность"³ -

1. Исследования по энтропии языка Р.А.Казаряна /см.Р.А.Казарян. Оценка энтропий армянского текста. "Известия АН Армянской ССР". Т.ХІУ, 1964, 4, стр.171/, а также Г.П. Богуславской /см. Г.П. Богуславская. Энтропия английского печатного текста. Канд.диссертация. Минск, 1966, стр. 44-45/ показали, что для достоверной оценки величины энтропии выборки указанного объема уже достаточны.

2. В.Г.Адмони. Еще раз об изучении количественной стороны грамматических явлений. ВЯ, 1970, № 1, стр. 95.

3. М.П.Кульгав. Синтаксические средства реализации сжатости / краткости / в современной немецкой научно-технической речи. Ученые записки. М., 1963 / I-й Московский государственный педагогический институт иностранных языков/, стр. 319.

Таблица I

Распределение энтропии и избыточности в немецком тексте длиной в 200 букв

код буквы	язык в целом						устно-разговорный						беллетристический						публицистический						научно-технический							
	н	н̄	р	ṛ	н	н̄	н	н̄	р	ṛ	н	н̄	р	ṛ	н	н̄	р	ṛ	н	н̄	р	ṛ	н	н̄	р	ṛ	н	н̄	р	ṛ		
30	0,80	1,57	67,0	83,2	0,96	1,87	60,7	79,8	0,98	1,62	66,0	79,4	0,65	1,32	72,3	86,4	0,41	0,99	79,2	91,2												
40	0,68	1,33	72,1	85,7	0,58	1,11	76,7	87,8	0,77	1,37	71,2	83,8	0,69	1,16	75,4	85,4	0,45	0,90	81,1	90,5												
50	0,61	1,28	73,1	87,2	0,57	1,05	77,9	88,0	0,59	1,16	75,4	87,6	0,83	1,39	71,0	82,6	0,56	0,97	79,6	88,2												
75	0,67	1,22	74,6	85,9	0,58	1,06	77,7	87,8	0,88	1,47	69,1	81,5	0,58	1,01	78,8	87,8	0,43	0,84	86,6	91,0												
100	0,80	1,40	70,6	83,2	0,58	1,08	77,3	87,8	0,94	1,34	71,8	80,1	0,79	1,28	72,7	83,4	0,65	1,11	76,7	86,3												
150	0,79	1,34	71,9	83,4	0,74	1,22	74,6	84,5	0,95	1,39	71,7	80,0	0,72	1,23	74,2	84,9	0,65	1,08	77,3	86,3												
200	0,64	1,20	74,8	86,6	0,80	1,22	74,6	83,2	0,73	1,15	75,8	84,7	0,60	0,97	79,6	87,4	0,58	0,96	79,7	87,8												
Среднее значение	0,71	1,36	71,4	85,1	0,69	1,23	74,2	85,5	0,83	1,36	71,4	82,5	0,69	1,19	74,8	85,4	0,53	0,98	79,4	88,8												
н	- максимальное значение энтропии ¹																															
н̄	- минимальное значение энтропии																															
р	- максимальное значение избыточности																															
ṛ	- минимальное значение избыточности																															

H — максимальное значение энтропии¹

H̄ — минимальное значение энтропии

R — максимальное значение избыточности

R̄ — минимальное значение избыточности

¹ О применении математического аппарата для определения указанных величин см. работу Р.Г. Пиотровского "Информационные измерения печатного текста в сб.: "Энтропия языка и статистика речи". Минск, 1966, стр. 11-53.

в максимальной степени накладывают ограничения на возможность свободно связывать элементы между собой в данном стиле.

Низкая избыточность беллетристического стиля является результатом большей по сравнению с другими стилями неопределенности в выборе языковых элементов, поскольку здесь используются языковые средства, присущие обычно разнообразным стилям. Кроме того, наличие большого количества неожиданных сочетаний слов/различные стилистические приемы/, широкий круг лексики также увеличивает неопределенность в выборе лингвистических единиц.

При сравнении информационных характеристик трех стилистических разновидностей немецкого, английского, французского, русского и польского языков / см. табл. 2 / обнаруживается одна и та же тенденция к увеличению избыточности научного делового стиля. Менее избыточным является устно-разговорный стиль, представленный в письменной форме. В информационных оценках беллетристического стиля наблюдаются колебания. Так, оценка избыточности немецких и польских текстов беллетристического стиля ниже избыточности текстов устно-разговорного стиля. Аналогичная картина отмечается и в румынском языке. По данным Л.А. Новак в устно-разговорном стиле $R = 74,2\%$, $\bar{R} = 85,4\%$; в беллетристическом - $R = 73,8\%$, $\bar{R} = 83,8\%$.^I

I. Оценки избыточности стилистических разновидностей румынского языка взяты из работы Г.П. Богуславской, Л.А. Новак. "Энтропия английского и румынского языков". Минск, 1968, стр. 13

Таблица 2

Сравнительные данные об энтропии и избыточности пяти языков¹

Функциональный стиль	Немецкий язык			Польский язык			Русский язык			Английский язык			Французский язык		
	$\bar{H}$	H	R	$\bar{H}$	H	R	$\bar{H}$	H	R	$\bar{H}$	H	R	$\bar{H}$	H	R
Устно-разговорный стиль															
И	1,51	0,78	72,4	83,6	1,18	0,69	76,3	86,3	1,40	0,83	72,0	83,4	1,47	0,90	69,4
Э															
И															
Научно-деловой стиль															
И	0,94	0,49	80,1	89,7	0,83	0,53	83,6	89,5	0,83	0,49	83,4	90,1	0,82	0,37	82,9
Э															
И															
Беллетристика															
И	1,43	0,98	70,0	79,4	1,29	0,83	74,5	83,6	1,19	0,70	76,3	86,0	1,10	0,65	77,1
Э															
И															
Язык в целом															
И	1,43	0,82	70,0	82,8	1,28	0,76	74,7	85,0	1,37	0,82	72,1	83,6	1,35	0,74	71,9
Э															
И															

¹ Оценки максимальной $\bar{H}$ и минимальной $\bar{H}$ энтропии, максимальной R и минимальной R русско-, английско-, французско- и польско-язычных взяты из работы Р.Г. Пиотровского "Информационные измерения языка", стр. 59.

Значения $\underline{H}$, $\bar{H}$, $\underline{R}$, $\bar{R}$, представленные в таблице 2, показывают большое сходство в общих оценках энтропии и избыточности пяти языков. Данные об избыточности немецкого языка подтверждают еще раз предположение Р.Г. Пиотровского о том, что " избыточность развитого языка находится в пределах между 70 и 85%^I."

Для оценки достоверности полученных количественных характеристик указанных стилистических разновидностей немецкого языка был проведен контрольный эксперимент по угадыванию, затем с помощью математических критериев /среднего квадратического отклонения σ и коэффициента вариации W / были оценены расхождения в количественных показателях основного и контрольного эксперимента / см.табл.3/. Эти расхождения оказались незначительными, что позволяет нам считать полученные оценки энтропии и избыточности репрезентативными информационными характеристиками немецкой письменной речи.

Таблица 3

Сравнение результатов основного и контрольного экспериментов с помощью критериев σ , W ²

Функциональный стиль	σ		W		в %
	$\underline{H}$	$\bar{H}$	$\underline{H}$	$\bar{H}$	$\bar{H}$
Язык в целом	0,15	0,15	18,5		10,6
Беллетристика	0,22	0,24	23,8		16,7
Публицистика	0,30	0,20	37,2		15,4

1. Р.Г.Пиотровский. Информационные измерения языка, стр.58

2. Оценки энтропии и избыточности научно-делового и устного-разговорного стилей в основном и контрольном экспериментах почти совпали по абсолютным величинам/см.табл. 2, стр. 7 /, естественно, что применение названных критериев уже не понадобилось.

Анализ распределения статистической информации в тексте показывает, что перепады максимумов и минимумов энтропии наблюдаются на определенных буквенных позициях, что придает распределению информации в тексте волнообразный характер.

В связи с этим возник вопрос, какова лингвистическая природа такого распределения информации? С этой целью необходимо было определить информационную насыщенность гласных, согласных, пробелов, достоверных продолжений. Прибегнув к методу корреляционного анализа, мы получили следующие численные величины коэффициента корреляции между количеством энтропии на букву и выше названными переменными:

а/ по достоверным продолжениям	- 0,65
б/ по гласным	- 0,37
в/ по согласным	- 0,25
г/ по пробелам	- 0,16

Проведенное исследование дает основание считать, что волнообразное распределение информации в немецком тексте мало обусловлено появлением таких фигур, как гласные, согласные, пробелы.

Для выявления распределения статистической информации не только внутри текста, но и внутри его отдельных единиц была построена усредненная информационная схема текста^I
/ см.рис. I/.

I. О процедуре построения лексической схемы текста см.:
Р.Г.Пiotровский. Информационные измерения языка,
стр. 69-73;

G.Boguslavakaja, T.Koženec, P.Piotrovskij. Informational Estimates of Text. "Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung". Bd.24, Heft, 6, 1971, S.465-470.

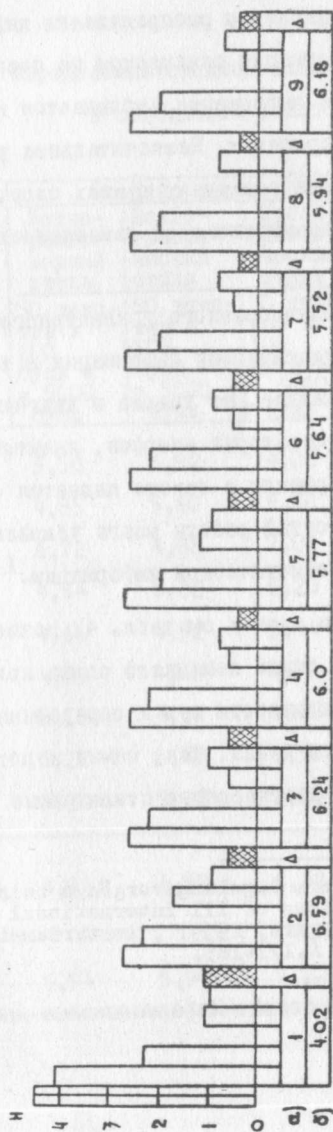


Рис.1. Информационная лексическая схема немецкого текста.

Условные обозначения: а/ номер слова в обобщенном тексте;

б/ средняя длина немецкого слова в буквах соответственно занимаемой позиции слова в обобщенном тексте.

Лексическая схема текста также обнаруживает волнообразный / квантовый/ характер распределения информации, а именно: максимумы информации приходятся на первые буквы слов, падение количества информации наблюдается на предпоследних буквах слов и на пробелах. Незначительное увеличение количества информации на последних буквах слов, по всей видимости, связано с грамматической изменяемостью большинства немецких слов.

Итак, причину волнообразного / квантового/ характера распределения статистической информации в тексте, слова следует искать, очевидно, не только в лингвистических факторах. По мнению некоторых авторов, квантовый характер распределения информации в тексте является своеобразным отражением особенностей работы мозга человека в процессе переработки им лингвистической информации.¹

Определение длины слов / см.табл. 4/ позволяет сделать вывод, что средняя длина немецкого слова колеблется в зависимости от принадлежности его к определенной стилистической разновидности языка. Так, самые короткие слова характерны для устно-разговорного стиля, самые длинные -

1. См.: K.Küpfmüller. *Regulungsvorgänge beim Sprechen und Hören. "Proceedings of III International Congress of Acoustics (Stuttgart, 1959)"*, Amsterdam-London-New York-Princeton, 1962, S.171-183.

Ср. также: Р.Г.Пиотровский. Информационные измерения языка, стр.103.

Таблица 4

Определение средней длины слова
в обобщенном тексте

Объем выборки	Средняя длина слова в буквах					Язык в целом
	Устно- разго- ворный стиль	Белле- тристи- ческий стиль	Публи- цисти- ческий стиль	Научно- техни- ческий стиль		
	50 текс- тов х 100 букв = 5000	50 текс- тов х 100 букв = 5000	50 текс- тов х 100 букв = 5000	50 текс- тов х 100 букв = 5000	200 текс- тов х 100 букв = 20000	
	букв	букв	букв	букв	букв	
№ сло- ва в тексте	1	3,58	3,92	4,00	4,56	4,02
	2	4,85	5,60	6,68	9,22	6,59
	3	6,72	4,88	7,08	6,26	6,24
	4	4,54	5,22	6,20	8,06	6,00
	5	4,66	4,92	5,44	8,06	5,77
	6	4,58	4,66	6,72	6,60	5,64
	7	5,10	4,80	6,18	6,00	5,52
	8	5,22	5,48	6,50	6,54	5,94
	9	5,24	5,22	7,14	7,12	6,18
Средняя длина слова по сти- лям	Без пробела	4,94	4,97	6,22	6,94	5,76
	С про- белом	5,94	5,97	7,22	7,94	6,76

для научно-технического.¹

Определение длины предложения путем подсчета количества слов, входящих в него, позволяет, вслед за Г.Аренсом, говорить о наличии в современном немецком языке тенденции к взаимозависимости длины предложений и длины слов.

Исследование синтактики современного немецкого текста с помощью экспериментального метода Шеннона-Пиотровского позволило получить сведения об информационно-статистической структуре немецкого текста и его четырех функциональных стилей.

Полученные данные позволяют сделать следующие наблюдения:

1. Сопоставление информационных оценок по стилям немецкого языка показало, что наименьшей энтропией и наибольшей избыточностью характеризуется научно-технический стиль. Наиболее информационно-насыщенным является беллетристический стиль. Информационная оценка с этой точки зрения устно-разговорного стиля несколько ниже. Теоретико-информационные характеристики публицистического стиля близки к аналогичным характеристикам научно-технического стиля. Оценка избыточности языка в целом приближается к оценкам избыточности устно-разговорного и беллетристического сти-

1. Ср.: H. Arens. *Verborgene Ordnung. Die Beziehungen zwischen Satzlänge und Wortlänge in deutscher Erzählprosa vom Barock bis heute*. Düsseldorf, 1965, S. 55, 56, 85, 88;

В.Д. Давкин. Особенности немецкой разговорной речи. М., 1965, стр. 9; Э.Г. Ризель. Полярные стилистические черты и их языковые воплощения. "Ин. яз. в школе", 1961, № 3, стр. 98; И.М. Яглом, Р.Л. Добрушин, А.М. Яглом. Теория информации и лингвистика. ВЯ, 1960, 4, стр. 100.

лей. Она хорошо согласуется с оценкой избыточности, полученной К.Кюпфмюллером^I при помощи метода, основанного на статистической обработке частотных словарей.

2. Информационные оценки немецкой письменной речи близки к значениям этих величин в других языках.

3. Лексическая схема немецкого текста обнаруживает квантовое распределение информации. Кванты статистической информации сосредоточены здесь на начальных буквах слов. Последние буквы слов и пробел характеризуются малой информационной насыщенностью. "Зернистое" распределение статистической информации в тексте объясняется, очевидно, тем, что воспроизведение и переработка текста соответствующими центрами мозга происходит путем периодической отдачи квантов информации. Одновременно построенная нами модель текста показывает, что статистические связи внутри слов текста сильнее, чем контекстные связи между словами.

Вторая глава

Здесь рассматриваются вопросы, связанные с информационной структурой немецкого слова на буквенном, ологовом /силлабографном/морфемном /морфографном/уровнях и делительного сложного слова на уровне составляющих его

1. По данным К.Кюпфмюллера оценка избыточности немецкого языка равна 69% / см. К.Кюпфмюллер. Die Entropie der deutschen Sprache. "Fernmeldetechnische Zeitschrift" VII, 6, 1954, S.240. /

основ. Анализируется характер распределения информации в текстовом и внетекстовом слове. Исследуется влияние лексико-грамматического контекста на информационную структуру слова. Измеряется грамматическая информация, содержащаяся во флексиях, а также смысловая информация некоторых элементов текста.

Решение поставленных задач требует рассмотрения вопросов, связанных со слоговым и морфемным членением немецкого слова.

Деление исследуемого материала на слоги проводилось по словарю Г. Варига. Что же касается морфемного членения, то оно специально исследовалось в одной из работ автора¹. Вся процедура морфемного членения словоформ состоит из следующей последовательности операций:

1. Выявление корневой морфемы.

2. Соотнесение оставшейся части словоформы со списком аффиксов современного немецкого языка.

3. Выявление нулевой морфемы.

Для выявления побуквенного, слогового и морфемного распределения информации в немецком текстовом и внетекстовом слове исследованию подвергалась выборка объемом 1565 немецких слов длиной от трех букв / две буквы плюс пробел/

1. См. Т.Ф. Кожанец. Информационная структура немецкого слова. "Частные вопросы автоматического анализа текста". Минск, 1972, стр. 35-50.

до семнадцати букв / шестнадцать букв плюс пробел/.

1015 слов общей выборки отгадывались при условии, что информанту известен предшествующий контекст длиной в 30 букв / включая пробел /¹. Это так называемые текстовые слова.

Остальные 550 слов выборки угадывались вне контекста /будем называть их внетекстовыми/. Для анализа исследования распределения информации в определительных сложных словах послужило 150 слов.

Полученные результаты представлены в виде распределения информации по буквам / рис.2/, по слогам / рис.3, 3 а/, по морфемам / рис.4, 4 а/ и по основам / рис.5, 5а: 5б/. Построение обобщающих схем на буквенном уровне ограничено шестнадцатой буквой, поскольку слова длиной более шестнадцати букв в нашей выборке составили около 5%. Построение слоговых и морфемных схем ограничено длиной слова в шесть слогов и в шесть морфем.² Информационные схемы определительных сложных слов охватывали словоформы из двух, трех и четырех основ.³

1. См.: N.G.Burton, J.C.Licklider. Long-Range Constraints in the Statistical Structure of Printed English. "The American Journal of Psychology". Vol. 68, N 4, 1955 pp. 650-653.

2. См.: В.А. Москович, Лексико-семантическая система естественного языка и информационные языки. Автореферат доктор. диссертации. М., 1971, стр. 42; H.Arens.

3. См.: J.Grimm. Deutsche Grammatik. Zweiter Teil. Neuer vermehrter Abdruck. Besorgt durch W.Scherer. Gütersloh (o.J.), S. 92.

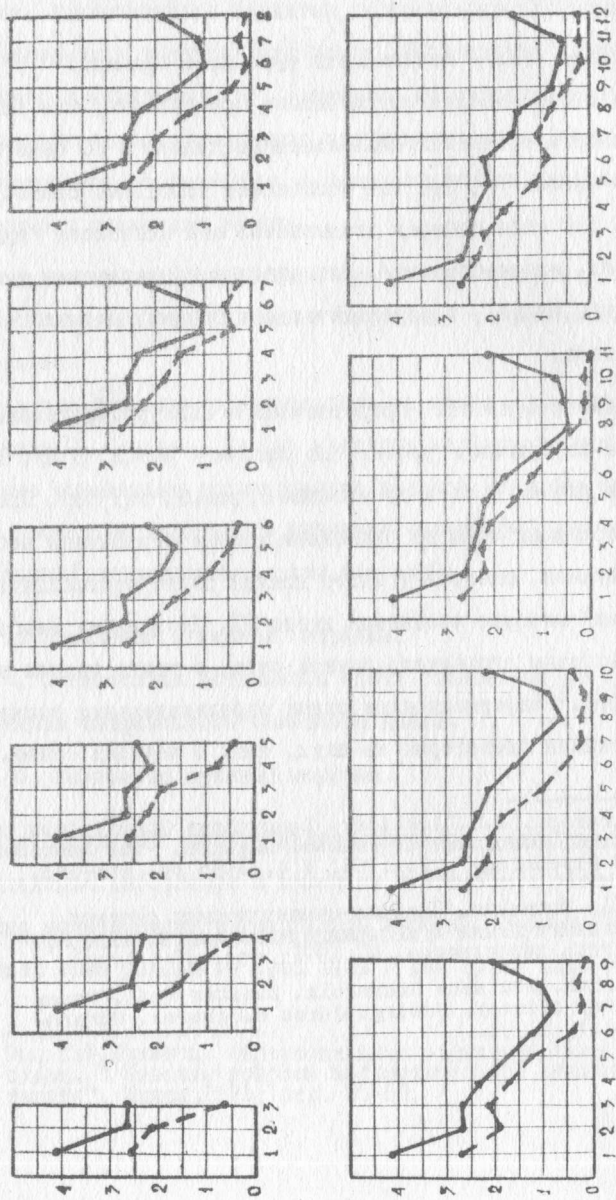


Рис. 2. Подкупленное распределение информации в текстовых источниках (штриховая линия) и во вне-
текстовых источниках (сплошная линия).

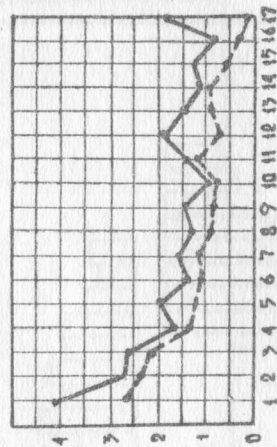
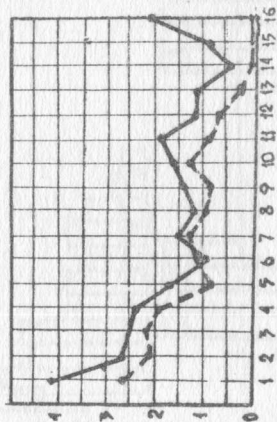
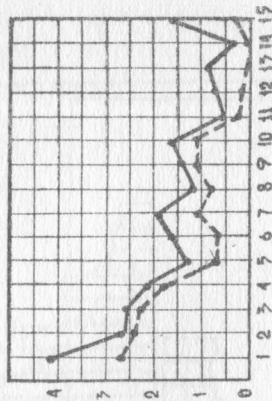
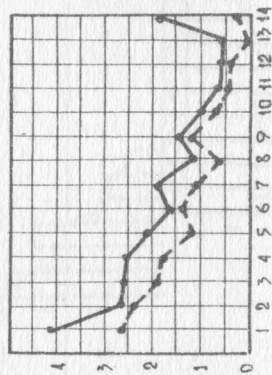
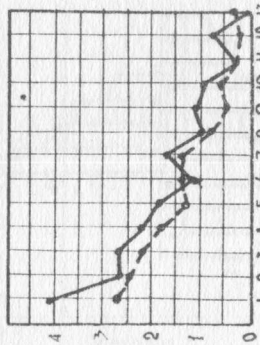


Рис.2. (продолжение) .

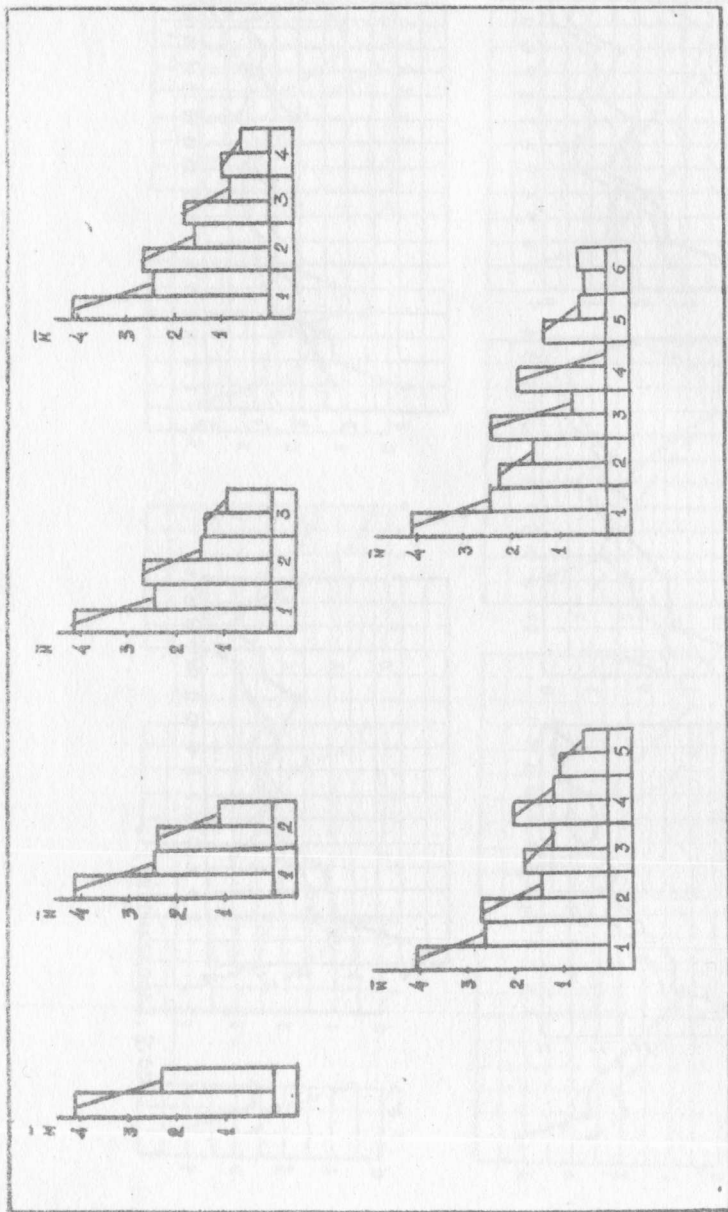


Рис. 3. Слововое распределение информации во внетекстовых словах.

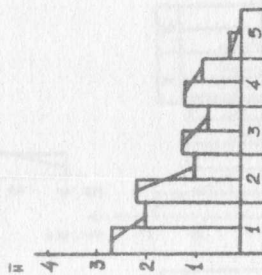
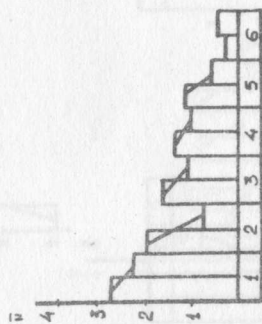
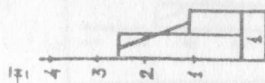
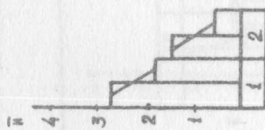
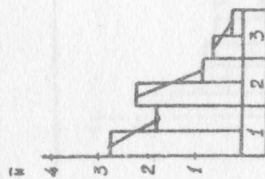
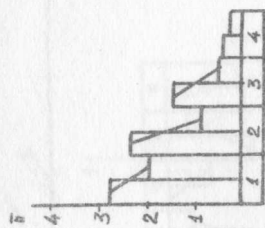


Рис. 3а. Словговое распределение информации в текстовых словах.

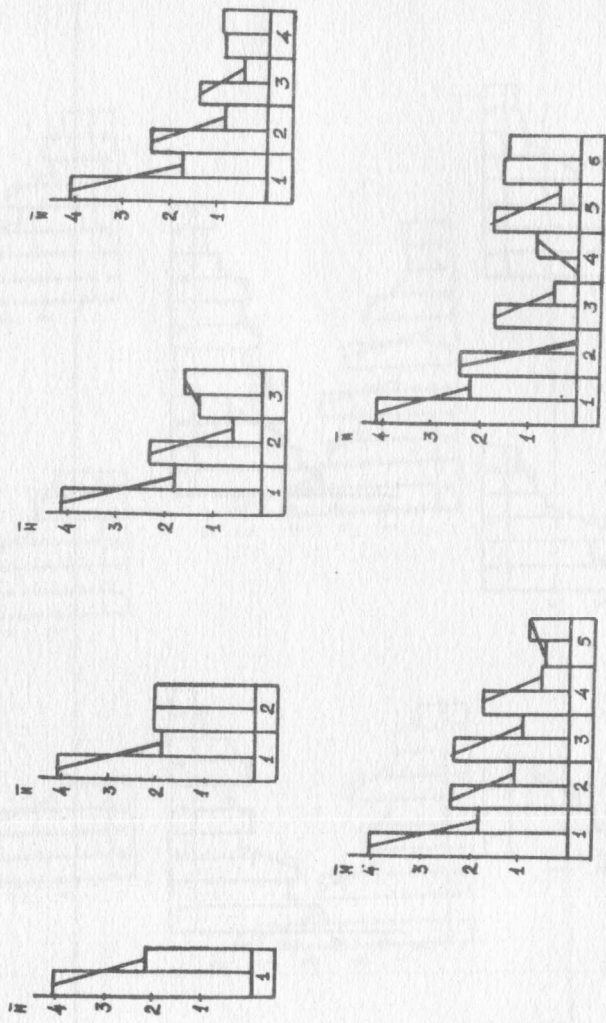


Рис. 4. Морфемное распределение информации во внятекстовых словах.

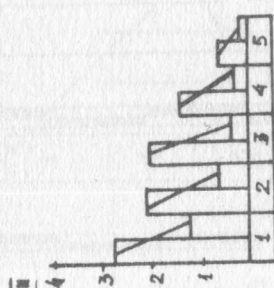
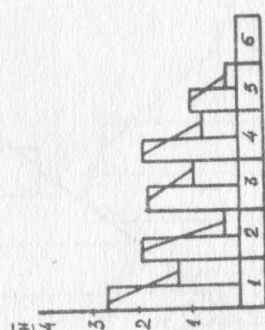
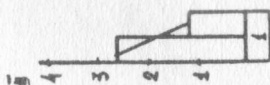
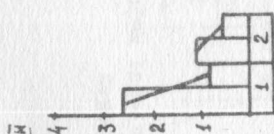
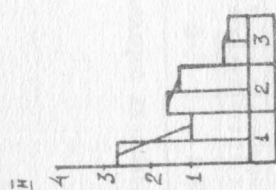
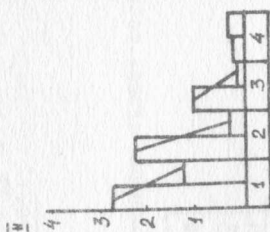


Рис. 4а. Морфемное распределение информации в немецком текстовом слове.

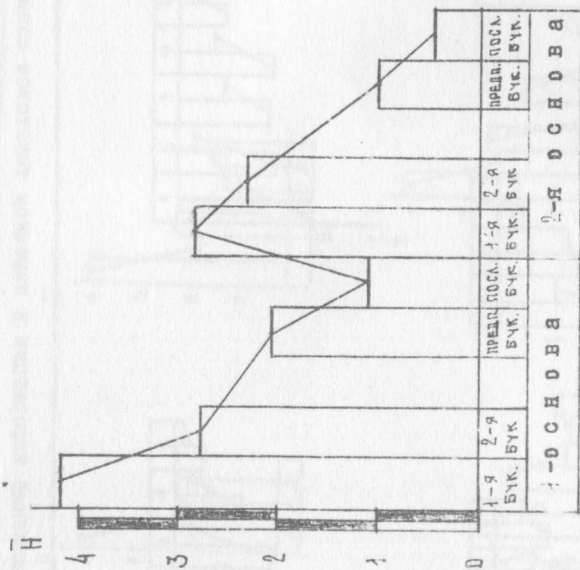


Рис.5. Распределение информации в немецких определительных словах, состоящих из 2-х основ

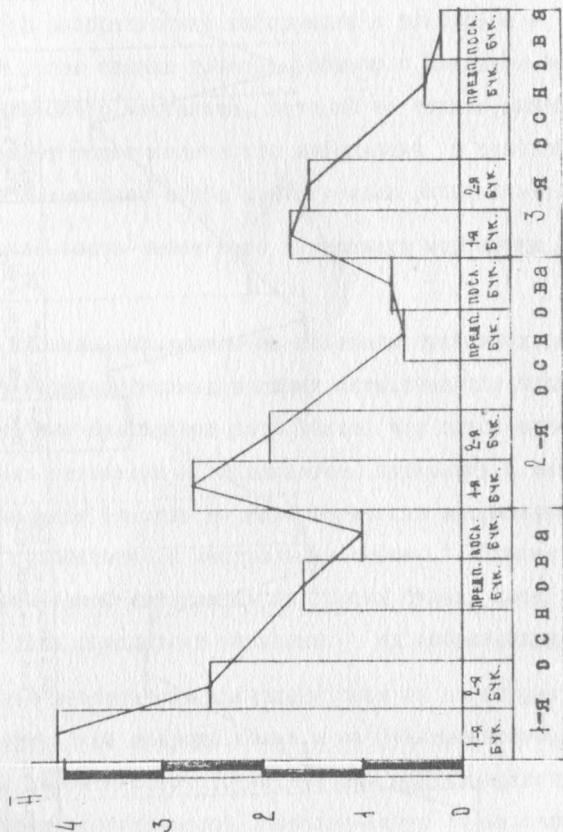


Рис.5а. Распределение информации в нематриках определительных сложных словах, состоящих из 3-х основ.

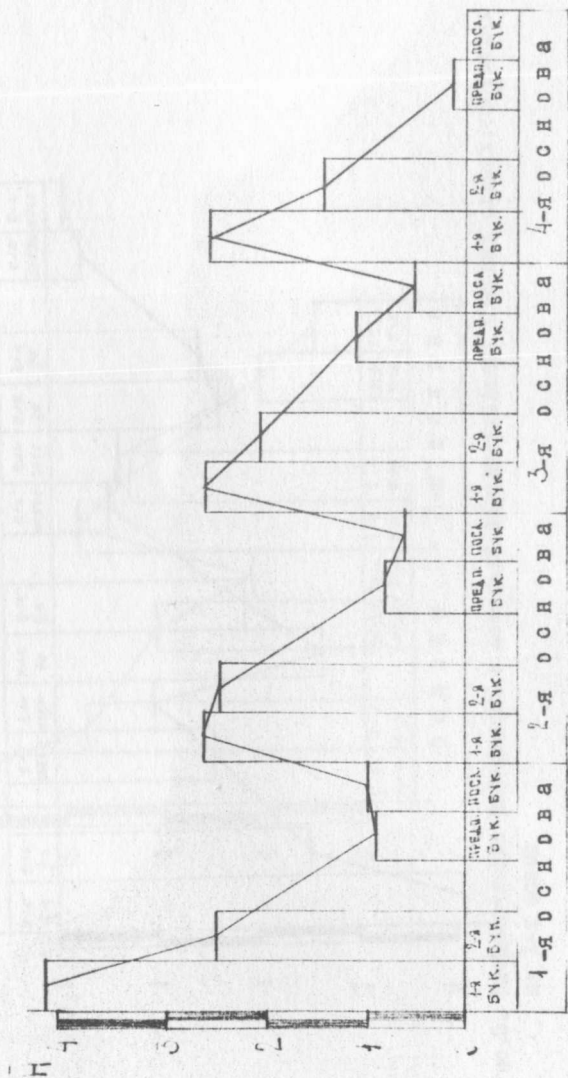


Рис.5б. Распределение информации в немецких определительных сложных словах, состоящих из 4-х основ.

Осуществленное исследование позволило сделать следующие выводы:

1. Характер распределения информации в текстовом и внетекстовом слове связан главным образом с влиянием лексико-грамматического контекста, который не только значительно уменьшает общее количество информации у текстового слова, но и изменяет общую конфигурацию распределения.

2. Немецкое слово имеет ярко выраженную морфемную структуру.

3. Резкое падение информации на конечных буквенных позициях основ определительных сложных слов, очевидно, можно рассматривать как проявление того факта, что здесь никаких грамматических отношений не выражается, поскольку в словосложении принимают участие не морфологически оформленные элементы, а грамматически нейтральные основы.^I Концентрация статистической информации на первых буквах всех основ сложных слов определяет механизм их аббревиации.

4. Все эти особенности проливают свет на некоторые аспекты семиотической природы языка и ее функционирования в речи. Так, вероятностно-статистические закономерности, характеризующие сочетаемость символов-фигур /букв, слогов/, действует на коротких начальных участках текста. Как только следующие друг за другом по синтагматической оси

I. См.: К.А.Левковская. Лексикология немецкого языка. М., 1956, стр.180; Duden. Grammatik der deutschen Gegenwartssprache. Bd.4, Leningrad, 1962, S.330.

фигуры формулируют знак / морфему, основу/, начинают действовать вероятностно-статистические закономерности сочетаемости знаков и т.д. Комбинации элементарных символов-фигур подчиняются правилам комбинаторики более сложных символов /знаков/ и в целом определяются законами языка, спецификой его строя, звуковыми, орфографическими и лексико-грамматическими ограничениями.¹

Исследование длины немецкого слога в зависимости от длины слова подтверждает высказанное П.Менцератом предположение о наблюдаемой в немецком языке тенденции количественного соотношения между числом слогов в слове и числом приходящихся на каждый силлабограф букв.² Так; в словах, имеющих только один слог, средняя длина слога равна 3,5 буквы, в словах с шестью слогами она равна 2,6 буквы.

Нами установлено, что наблюдается такая же аналогичная картина количественного отношения между числом морфем в слове и количеством букв в морфографе. В одноморфемных словах средняя длина морфемы равна 4,5 буквы, в шестиморфемном - 2,7 буквы.

Исследование буквенного и слогового наполнения определенных сложных слов подтверждает также высказывание П.Менцерата о том, что чем длиннее слова / т.е. чем больше слогов они содержат, тем меньше радиус колебаний в них числа звуков.³

1. Ср. Р.Г.Пиотровский. Информационные измерения языка, стр.86-89.

2. P.Menzerath. Die Architektur des deutschen Wortschatzes. "Phonetische Studien". Heft 3, Bonn, 1954, S.100.

3. Там же. S. 103.

Так, в нашей экспериментальной выборке средняя длина слога 2-х, 3-х и 4-х основных слов равна $\sim 3,3$ буквы.

Нами установлено, что информация, сообщаемая словом, / внутрисловные ограничения/ выше более чем в два раза информации, передаваемой лексико-грамматическим контекстом / контекстные ограничения/. Как показывают данные таблицы 5, в аналитическом английском языке контекстные ограничения составляют более $1/3$ всех ограничений, накладываемых на слово.¹

Таблица 5

Язык	f в %
Английский	35,0
Немецкий	29,2
Французский	23,5
Русский	22,6

Меньшая оценка контекстных ограничений для немецкого языка в сравнении с английским, по всей видимости, отражает тот факт, что немецкий язык, будучи одновременно языком аналитическим и флективным, характеризуется несколько большей свободой сочетаемости слов. Ясно, что

1. О методике расчета оценки информации, передаваемой лексико-грамматическим контекстом см. Т.Ф. Коженец. О количественных оценках лексико-грамматических связей слова в немецком печатном тексте. "Вопросы романогерманской и славянской филологии". Минск, 1971, стр. 117-120.

"сочетание слов, характеризованное окончанием, является менее связанным, чем сочетание, части которого за отсутствием внутренней характеристики опираются друг на друга".¹

Получив количественную оценку лексико-грамматической информации, снимаемой контекстом, мы смогли определить долю содержащейся в ней грамматической /флексивной/ информации. В данном случае речь идет об определении количественной оценки информации следующих морфем -окончаний: -е, -(е)в, -(е)н, -ер, -енс, -st и нулевой флексии.

Математические расчеты показали, что доля грамматической информации / флексии/ в общей сумме лексико-грамматической информации, снимаемой контекстом, равна 10,81%.²

Причина относительно невысокой оценки флексивной морфологии в немецком языке, по-видимому, объясняется наличием в нем большего количества грамматических плеоназмов, средств замены флексии / например, артикль, предлоги, местоимения, вспомогательные глаголы и т.д./, омографичностью флексии и т.д. Иными словами, наличие в немецком языке служебных слов, функционально эквивалентных флексии, с одной стороны, и повторение грамматических значений как в согласованных членах предложения, так и в пределах форм

1. Ш.Балли. Общая лингвистика и вопросы французского языка. Перевод с франц. М., 1955, стр.314

2. Подробное описание методики данного расчета см.: Н.В.Петрова, Р.Г.Пиотровский. Слово, контекст, морфология. Вп, 1966, 2, стр. 119-122

одного и того же слова, с другой стороны, обуславливает относительно невысокий информационный вес немецкой флексии. Естественно предположить, что внутренняя флексия в какой-то мере уменьшает информационный вес немецких окончаний / этот вопрос еще исследуется нами/

x

x

x

Различные приемы психолингвистического эксперимента Шеннона-Питровского дали возможность получить численные оценки синтактической / селективной/ информации, содержащейся в различных отрезках текста. Как уже отмечалось, эта методика основана на предположении о том, что в постоянной памяти носителя языка сохраняются вероятностные и условно-вероятностные характеристики отдельных лингвистических единиц и связей между ними в тексте.

Исходя из этих предположений, можно построить методику измерения семантико-прагматической информации, опирающуюся на оценки синтактической информации.

Пусть имеется две оценки энтропии / неопределенности, количественно равной синтактической информации/ сегмента ℓ^n . Одна - H / ℓ^n / получена так, что предшествующий участок ℓ^{n-1} , состоящий из одного или нескольких слов, неизвестен идеальному угадчику. Вторая - H' / ℓ^n / получена при условии, что предстория ℓ^{n-1} известна угадчику. Совершенно естественно, что неопределенность H / ℓ^n

больше энтропии $H' / \mathcal{B}^n$ / на величину семантико-прагматической / смысловой / информации, полученной угадчиком из участка $\mathcal{B}^{n-1}$. Итак, семантико-прагматическая информация слова / или словосочетания /, составляющего участок $\mathcal{B}^{n-1}$, оценивается на диагностирующем участке $\mathcal{B}^n$ с помощью разности

$$I(\mathcal{B}^{n-1}) = H(\mathcal{B}^n) - H'(\mathcal{B}^n)$$

Полную оценку смысловой информации $I / \mathcal{B}^{n-1}$ / можно получить при условии, что сегмент $\mathcal{B}^n$ достаточно велик, чтобы в нем реализовались все правые валентности слова или словосочетания $\mathcal{B}^{n-1}$. Так, например, в нашем эксперименте использовались цепочки: Der Krieg mußte bald und schlecht enden. и Der Mann lümmete sich über seine Koffer. в которых семантика начальных словосочетаний / der Krieg, der Mann / измерялась путем сравнения результатов двойного угадывания следующих за ними контекстами.

Вместе с тем нельзя не учитывать того, что оценка $I / \mathcal{B}^{n-1}$ / может меняться в зависимости от того, каким диагностирующим контекстом $\mathcal{B}^n$ мы пользуемся. Чтобы уменьшить влияние диагностирующего контекста на величину семантико-прагматической информации, будем использовать для ее оценки также смысловой коэффициент, выражаемый отношением:

$$S = \frac{I(\mathcal{B}^{n-1})}{H(\mathcal{B}^n)}$$

По описанной методике измерялась смысловая информация словосочетаний *der Krieg* и *der Mann* в ста немецких простых распространенных предложениях с прямым порядком слов длиной в 30 букв. В эксперименте, осуществленном по индивидуальной программе угадывания, участвовало 8 студентов немецкой национальности из Белорусского политехнического института г. Минска.

Результаты исследования представлены в таблице 6

Таблица 6.

№ п/п	Контроль- ное слово	$H(\ell^n)$	$H'(\ell^n)$	$H(\ell^{n-1})$	$S = \frac{H(\ell^{n-1})}{H(\ell^n)}$
		в дв.ед.	в дв.ед.	в дв.ед.	
1.	<i>der Krieg</i>	50,28	45,23	5,05	0,10
2.	<i>der Mann</i>	52,54	50,31	2,23	0,04

Наблюдения над полученными количественными характеристиками смысловой информации позволяют сделать следующие предварительные заключения:

I. Величина смысловой информации не коррелирована жестко со значением синтаксической информации. Так, синтаксическая информация словосочетания *der Krieg* равна 50,28 дв.ед., смысловая - 5,05 дв.ед., словосочетания *der Mann* - соответственно 52,54 дв.ед. и 2,23 дв.ед. Это позволяет нам говорить о более высоком уровне лексической семантики словосочетания *der Krieg*, чем словосочетания *der Mann*, поскольку грамматические значения у них одинаковые.

2. Оценка смысловой информации лингвистических единиц, в отличие от оценки синтаксической информации, не находится в обратной зависимости от частоты употребления лингвистических единиц. По словарю Венглера общая частота употребления слова *Krieg* равна 50, слова *Mann* — 184^I. Частота употребления артикля здесь во внимание не принимается, так как она в обоих случаях одинакова и, естественно, оказывает одинаковое влияние на предсказуемость букв последующего текста.

Результаты статистического эксперимента могут послужить моделью для измерения семантико-прагматической информации в естественном языке, поскольку в ходе исследования языкового текста экспериментом по угадыванию оказывается практически невозможно разединить и рассмотреть в чистом виде его синтаксические и прагматические функции.²

Целесообразность исследования семантики текста методами теории информации определяется прежде всего практическими потребностями: повышением пропускной способности каналов связи посредством оптимального кодирования смысловой и прагматической информации; установлением относительно разных возрастных групп порога информативности текста,

I. См. Н.-Н. Wängler. *Rangwörterbuch hochdeutscher Umgangssprache*. Marburg, 1963, S. 39, 49.

2. Г.П. Богуславская, П.Б. Невельский. Рецензия на работу Р.Г. Пиотровского "Информационные измерения языка". ВЯ, 1970, № 3, стр. 134. Ср. также: В.А. Звегинцев, *Применение в лингвистике логико-математических методов*. Сб. "Новое в лингвистике", вып. IV, М., 1965, стр. 20, 21; К.Е. Морозов. *Философские проблемы теории информации*. Сб. "Философия естествознания", вып. I, М., 1966, стр. 383-408.

искаженного в условиях автоматической переработки; оценкой степени трудности текста как один из методов в преподавании иностранного и родного языка и др.^I

ЗАКЛЮЧЕНИЕ

Распределение и количество статистической информации в слове, тексте находятся в прямой зависимости от величины семантической и прагматической информации, что подтверждается информационным весом слов, морфем, обусловленным частотой их употребления в речи.

Получена невысокая информационная оценка морфологических аффиксов в немецком языке. Наличие во флективно-аналитическом немецком языке средств замены флексий, обилие грамматических плеоназмов, омографичность флексий и т.д. способствуют меньшему информационному весу флексии немецкого языка по сравнению с флексией аналитического французского и синтетического русского языков.

Распределение статистической информации в немецком определительном сложном существительном характеризуется ее высокой концентрацией на первых буквах и резким падением на последних буквах всех основ. Очевидно, при обработке немецких определительных сложных слов на ЭВМ они мог-

I. В.И.Богодиот, Х.П.Георгиев, В.Н.Пастунова, С.В.Райгар. Измерение семантической информации текста. Материалы Всесоюзной научной конференции 6-8 октября 1971 г., г. Кишинев/ стр. 261.

ли бы быть введены в машину в аббревированном виде. Это позволило бы значительно сократить количество ячеек, необходимых для записи в памяти ЭВМ слов текста. В связи с этим возникает проблема информационного исследования структуры немецкого сложного слова вообще с целью выявления его пограничных сигналов и т.д.

Факт неравномерного распределения статистической информации в слове может иметь принципиальное значение при расчете памяти анализирующей и синтезирующей текст ЭВМ.

Эксперимент по угадыванию описывает с помощью информационно-вероятностных измерений действие структурных связей и их иерархию в тексте.

Ценность метода угадывания состоит также в том, что он может применяться для оценки семантической информации, для объективной оценки степени овладения языком, для оценки трудности текстов в учебниках для того или иного состава учащихся и т.д.

Основные положения диссертации изложены в следующих опубликованных работах:

1. Об информационной характеристике морфемной структуры немецкого слова. "Актуальные проблемы лексикологии". Тезисы докладов и сообщений Всесоюзной научной конференции 17-20 июня 1970 г. г. Минск, 1970.

2. О количественных оценках лексико-грамматических связей слова в немецком тексте. "XXII научная сессия Минского

государственного педагогического института иностранных языков". Тезисы докладов. Часть II. Языкознание. Минск, 1970.

3. Информационное исследование немецкого текста. "Автоматическая переработка текста методами прикладной лингвистики". /Материалы Всесоюзной научной конференции 6-8 октября 1971 г., Кишинев/.

4. Информационная структура немецкого слова. "Частные вопросы автоматического анализа текстов". Минск, 1972.

5. О количественных оценках лексико-грамматических связей слова в немецком печатном тексте. "Вопросы романо-германской и славянской филологии". Минск, 1971.

6. Измерение смысловой информации. "Частные вопросы автоматического анализа текстов". Минск, 1972 / совместно с Х.П. Георгиевым и др./.

7. Informational Estimates of Text. "Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung" Bd. 24, Heft 6, 1971. / совместно с Г.Богуславской и др./.

8. Language and Information. "Biuletyn Fonograficzny", Poznań, XIII, 1972. /совместно с Г.Богуславской и др./.

1693

Библиотека
МГПИИЯ

Подписано к печати 2.Ш.1973 г.

Бумага 60 x 84 I/16. Печ.л.2,5. Уч.-изд.л.1,1.

Заказ I4. Тираж 250 экз. Бесплатно.

Отпечатано на роталпринте Минского государственного педагогического института иностранных языков.

Минск, ул.Захарова, 21.