АКАДЕМИЯ НАУК СССР
ИНСТИТУТ ЯЗЫКОЗНАНИЯ

Р. Г. ПИОТРОВСКИЙ

ИНФОРМАЦИОННЫЕ ИЗМЕРЕНИЯ ЯЗЫКА

ИЗДАТЕЛЬСТВО
«НАУКА»
ЛЕНИНГРАДСКОЕ
ОТДЕЛЕНИЕ
1968

В работе приводятся важнейшие статистические и информационные оценки для индоевропейских (русский, польский, английский, французский, румынский, испанский) и тюркских (азербайджанский, казахский) языков. Эти оценки (энтропия, избыточность, контекстная обусловленность, распределение информации в слове и словосочетании) могут быть использованы при автоматическом анализе и синтезе текста, а также при разработке эффективной методики преподавания языков. Работа рассчитана на языковедов и может быть использована в качестве учебного пособия для студентов и аспирантов, изучающих математическое языкознание.

7—1—1
188—1967 (I пол.)

Посвящается памяти поэтов:
М. ПИОТРОВСКОЙ, И. БЕСКЛУБОВА,
П. ГРУДИНИНА, Я. УРБАНЬСКОГО,
В. ШАНЯВСКОГО.

Глава I.

ЭКСПЕРИМЕНТ ПО УГАДЫВАНИЮ БУКВ НЕИЗВЕСТНОГО ТЕКСТА

§ 1. ИНФОРМАЦИОННЫЕ ИЗМЕРЕНИЯ ЕСТЕСТВЕННОГО ЯЗЫКА

Структура естественного языка в прямом наблюдении не дана. Элементы этой структуры и их связи не может вскрыть ни простое перечисление, ни описание наблюдаемых фактов, ни логистические рассуждения по поводу этих фактов. Структурные элементы и связи естественного языка могут быть вскрыты лишь путем применения различных экспериментов. Выбор конкретного эксперимента зависит от того, какая сфера языка подвергается исследованию, а также от того, каковы задачи этого исследования.

Поскольку естественный язык является важнейшим средством общения, следует предположить, что его единицы и связи могут быть обнаружены с помощью экспериментов и оценены путем измерений, использующихся в семиотике и теории информации.

Естественный язык, как и всякая иная, передающая информацию, семиотическая система, может рассматриваться в трех аспектах: синтактическом, семантическом и прагматическом. Синтактика рассматривает комбинаторику символов (знаков), отвлекаясь при этом от их значения и от той ценности, которую представляют эти символы и их сочетания как для лица или устройства, передающего информацию (передатчика), так и для потребителя информации (интерпретанта). Семантика изучает отношения между символами и их содержанием (или, в терминах семиотики, — между означающим и означаемым), игнорируя при этом как производителя, так и интерпретанта этих символов. Наконец, прагматика исследует символы с точки зрения их ценности для интерпретанта сообщения [66; 72, стр. 14;

78, стр. 167 и сл.], а иногда и для передатчика информации.

Вероятно, при изучении естественного языка наибольший интерес представляли бы прагматические оценки информации. Ведь как раз на этом уровне осуществляется во всей полноте реальный процесс речевого общения между людьми. Именно на этом уровне символы рассматриваются в определенных обстоятельствах при определенном окружении с привлечением проблем их ценности, полезности, узнавания и интерпретации [ср. 50, стр. 256]. Однако для расчета прагматической информации необходимо иметь вероятностные оценки тех ситуаций, в которых может оказаться интерпретант, принявший и правильно понявший сообщение (для определения ценности информации с точки зрения передатчика необходимо соответственно оценить ситуации, в которых он окажется после передачи сообщения). Сложность этой задачи очевидна, и не случайно поэтому, что мы пока еще не располагаем работающей процедурой, с помощью которой можно было бы получать количественные оценки прагматической информации, содержащейся в языковом тексте. Некоторые предложения по поводу такой процедуры см. в работах [9, 37, 40а, 64, 78], указанных в перечне.

Выдвигаемые Й. Бар-Хиллелом идеи измерения семантической информации [см. 42, 43, 48] опираются на предложенную Р. Карнапом [47] теорию логических вероятностей. В качестве семантического аналога информации здесь выступает некоторая функция числа всех возможных условий, при которых сообщение было бы истинным. Для реализации этой идеи авторы истолковывают понятие «возможные условия» с помощью предложенного еще Л. Витгенштейном понятия «описание» [75, стр. 83 и сл.; 79]. При этом число элементов множества описаний зависит от того языка, с помощью которого осуществляется это описание [ср. 78, стр. 169].

Теория Карнапа—Бар-Хиллела рассматривает и оценивает семантико-информационное содержание лишь простых суждений (декларативных предложений), вне всякой связи с прагматическими вопросами полезности, ценности, эвристичности информации для ее потребителя. Иными словами, семантические вопросы связи (для речевой коммуникации это вопросы лингвистической истинности или семасиологические вопросы) остаются за пределами семантической теории Карнапа—Бар-Хиллела. Если учесть также, что практически невозможно определить весь ряд альтернативных логических описаний и распределить на него меру вероятности, то станет ясным, что перспективы применения только что рассмотренной семантической теории для исследования значений в естественном языке весьма неопределенны.

Наиболее реалистичным является измерение синтактической информации, содержащейся в речи. Как известно, применяемая в этой процедуре мера информации основана на статистической частоте появления сигналов-символов при полном отвлечении от их значения и ценности (ср. выше). Исходя из этих особенностей, многие исследователи делают вывод о том, что мера Шеннона может быть применена только к статистически стационарному источнику сигналов [ср. 50, стр. 238—239]. Что же касается естественного языка, то, согласно указанной точке зрения, мера Шеннона может быть использована в лучшем случае лишь для статистического исследования структуры языка в том виде, в каком она существует в сознании того носителя языка, который выступает в роли источника сообщения. Эта мера, говорят они, ничего не дает ни с точки зрения оценки содержания, в том числе лингвистической ценности сообщения, ни с точки зрения отражения в этом сообщении объективной действительности, ни с точки зрения измерения того, как интерпретирует сообщение его получатель, ср. [83].

Само собой разумеется, что информационные оценки Шеннона не могут быть использованы для непосредственного измерения семантики и прагматики естественного языка. Однако нельзя забывать о том, что синтактика, семантика и прагматика естественного языка не просто находятся в тесном взаимодействии, но определенным образом коррелированы и взаимообусловлены. Каждое сообщение выбирается в теории Шеннона из некоторого множества возможных сообщений. Возможность же появления этих сообщений определяется, как правило, соотношением, существующим между языком и текстом, с одной стороны, и объективной реальностью — с другой (об особом характере этого соотношения см. в работах [62, стр. 59—60; 78, стр. 116—123]. Эта соотнесенность семантики (как в аспекте сигнификаций, так и в сфере референций, ср. [71]) и синтактики является, в частности,

условием для эвентуальных приложений статистической теории Шеннона к семантической теории Карнапа—Бар-Хиллела. Что же касается возможности перенесения результатов статистических экспериментов на прагматику естественного языка, то ее предпосылкой служит соотнесенность языка-кода и порождаемых им текстов с коллективным опытом носителей данного языка и коллективной оценкой ими пользы и эвристичности их содержания.

Из всего сказанного следует, что результаты статистического эксперимента могут послужить в будущем моделью для измерения семантической и прагматической информации в естественном языке.

§ 2. СТАТИСТИЧЕСКИЕ ИЗМЕРЕНИЯ ИНФОРМАЦИИ В РЕЧИ. ЭКСПЕРИМЕНТ ПО УГАДЫВАНИЮ БУКВ НЕИЗВЕСТНОГО ТЕКСТА

Статистические измерения информации в речи осуществляются обычно исходя из следующих предположений:

1) речевое общение рассматривается как канал связи, по которому с помощью букв, звуков, морфем или других лингвистических единиц передается информация (ср. рис. 1);

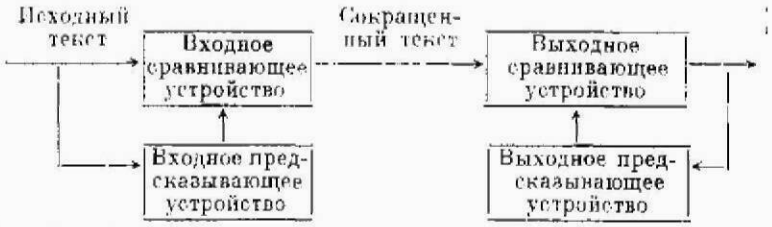

Рис. 1. Система связи, использующая сокращенный текст.

2) лингвистические единицы речи выступают в виде символов некоторого кода (языка); [1] в языке-коде заданы ограничения не только на сочетаемость (комбинаторику), но и на вероятность появления этих единиц в речи;

[1] В качестве кода выступает любая система символов при условии, что она по предварительному соглашению между отправителем и получателем используется для представления, хранения и передачи информации.

3) соединенные речевым каналом источник и приемник сообщения, т. е. говорящий и слушающий, используют один и тот же код.[2]

Последнее условие введено для того, чтобы отвлечься от таких ситуаций, когда передаваемое по речевому каналу сообщение либо воспринимается в приемном устройстве, исходя из иного языкового кода, чем тот, который был использован в источнике сообщения (ср. восприятие украинской речи носителем русского языка), либо расшифровывается в условиях неполного владения данным языковым кодом (ср. восприятие иностранного языка студентом или школьником, изучающим этот язык). Впрочем, интерес могут представлять и такие эксперименты, в которых третье условие не соблюдается: приемное устройство расшифровывает сообщение, либо исходя из неполного знания кода, либо опираясь на другой код (ср. ниже в § 27).

В языкознании используется два подхода к определению понятия «количество информации» — к о м б и н а т о р н ы й и в е р о я т н о с т н ы й. Введенный недавно А. Н. Колмогоровым третий алгоритмический подход, опирающийся на теорию рекурсивных функций [16, стр. 7 и сл.], в лингвистике пока не применялся.

При комбинаторном подходе предполагается, что переменное x способно принимать значения, принадлежащие конечному множеству X (в лингвистических терминах — алфавиту), которое состоит из S элементов (в нашем случае — букв, фонем, слогов, морфем, слов и т. д.). Тогда «энтропия» (неопределенность) переменного x будет равна

$$H(x) = \log S. \tag{1}$$

Определяя через i значение переменного x, снимаем эту энтропию и сообщаем «информацию» (I), количественно равную

$$I = \log S. \tag{2}$$

[2] Использование одного и того же кода не исключает возможности различного использования этого кода со стороны передатчика сообщения (говорящего) и его интерпретанта (слушающего). Ср. в этой связи идеи Ч. Хоккета о принципиальных различиях между «грамматикой для слушающего» и теми грамматическими механизмами, которыми пользуется говорящий при порождении текста [59, стр. 139 и сл.].

Если переменные $x_1, x_2, \ldots, x_n$ независимо пробегают множества, состоящие соответственно из $S_1, S_2, \ldots, S_n$ элементов, то тогда

$$H(x_1, x_2, \ldots, x_n) = H(x_1) + H(x_2) + \ldots + H(x_n).$$

Эксперимент, построенный на комбинаторном подходе, дает возможность оценить гибкость речи, т. е. определить степень разветвленности ее продолжения на определенном шаге текста [ср. 16, стр. 4]. Иными словами, формулы (1) и (2) позволяют определить энтропию (или соответственно количество информации), характеризующую как алфавит языка в целом, так и наборы тех лингвистических единиц, которые могут быть употреблены в данном участке текста. Энтропию алфавита мы будем передавать символом H_0, а для обозначения информации, получаемой путем снятия этой неопределенности, используем символ I_0.[3]

Выше уже говорилось, что система естественного языка приписывает каждому элементу буквенного, фонемного, слогового и т. д. алфавитов определенные вероятности (условные и безусловные). Поэтому более содержательные результаты информационных исследованиях можно получить, используя вероятностный подход. Так, например, имея распределение только безусловных вероятностей элементов, образующих тот или иной лингвистический алфавит $p_1, p_2, \ldots, p_S$, мы можем получить энтропию первого порядка (H_1), которая по своему численному значению ближе к средней условной энтропии данного лингвистического элемента текста (H_∞), чем значение энтропии, полученное при учете одной лишь комбинаторики. Согласно второй теореме Шеннона, значение H_1 и соответствующее ему количество информации определяется как

$$H_1 = I_1 = -\sum_{i=1}^{S} p_i \log p_i. \qquad (3)^4$$

[3] Подробный математический анализ возможностей комбинаторного подхода дает А. Н. Колмогоров [см. 16, стр. 3—5].

[4] В дальнейшем мы будем опускать равенства $H = I$ как сами собой разумеющиеся и ограничиваться указанием на одну из этих величин.

Доказательство этой теории и рассмотрение необходимых для этого допущений см. в работах [36, стр. 227—231; 38, стр. 3—20; 53, стр. 8—14; 77, стр. 390].

Задача усложняется, когда необходимо учесть и неравновероятность во взаимозависимости элементов алфавита. Здесь энтропия рассчитывается исходя из следующих соображений.

Обозначим некоторую цепочку лингвистических элементов $l_1, l_2, \ldots, l_{n-1}$ символом b^{n-1}. Цепочка $^{n-1}$ рассматривается в качестве случайного события, принимающего частный вид (значение) i. Предположим при этом, что $p(b_i^{n-1})$ всегда меньше единицы, т. е. грубо говоря, каждый лингвистический алфавит содержит более чем один элемент. Непосредственно за цепочкой $^{n-1}$ следует позиция l_n. Появление того или иного элемента в этой позиции также рассматривается в качестве случайной величины, принимающей значение $j_k (1 \leqslant k \leqslant S)$. Для каждого значения i, которое может принять b^{n-1}, имеется условная вероятность $p(j_{i,k} | b_i^{n-1})$ того, что l_n примет значение j_k.

Средняя условная энтропия H_n для позиции l_n будет получена в результате осреднения энтропии, сосчитанной по всем значениям b_i^{n-1}, с весами, соответствующими вероятностям цепочек $n-1$. Таким образом, имеем

$$H_n = -\sum_{b_i^{n-1}} p(b_i^{n-1}) \sum_{k=1}^{S} p(j_{i,k} | b_i^{n-1}) \log p(j_{i,k} | b_i^{n-1}). \qquad (4)$$

Эта величина показывает, какова в среднем неопределенность выбора лингвистического элемента n, когда известна цепочка $n-1$ [13, стр. 263].

Если взаимосвязи элементов распространяются как угодно далеко, то энтропия на один лингвистический элемент будет равна

$$H_\infty = \lim_{n \to \infty} H_n. \qquad (5)$$

Поскольку величины средней условной энтропии зависят от распределения вероятностей элементов на n-ом шаге текста и от вероятностей появления b_i^{n-1}, постольку эти величины могут быть определены из статистики многоэлементных сочетаний. Расчетная формула в этом случае имеет следующий вид:

$$H_n = H(j | b_i^{n-1}) = H(b_i^n) - H(b_i^{n-1}). \qquad (6)$$

Объем работы при вычислении H_I, H_{II}, H_{III}, H_{IV} для буквенного и фонемного алфавитов сравнительно невелик. Существуют также реалистичные способы оценки H_I для слов, словоформ и даже словосочетаний [3, стр. 179 и сл.].

Однако число комбинаций из пяти, шести и т. д. букв (фонем), не говоря уже о комбинациях из четырех, пяти и т. д. слов (слогов, морфем) растет так быстро, что для определения условной энтропии этих и более высоких порядков даже при использовании упрощенных методов расчета [35, стр. 112; 67, стр. 114—125] требуется колоссальная счетная работа. Эту работу нелегко выполнить, даже используя вычислительную технику, возможности которой в смысле обработки больших массивов лингвистической информации весьма ограничены [ср. 14, стр. 108 и сл.]. Поэтому приходится искать какие-то косвенные приемы, с помощью которых было бы можно быстро оценить энтропию и информацию в разных участках текста.

Этим условиям отвечает эксперимент по угадыванию букв [5] *неизвестного текста с последующей математической* обработкой его результатов. В основе этого метода положены следующие соображения.

1. Эффективное угадывание возможно лишь при условии, что в тексте имеется избыточная информация.

2. Лингвистическое сознание носителя языка, угадывающего текст, рассматривается в качестве некоторого кибернетического устройства.

3. Физиологический механизм этого устройства в деталях неизвестен (оно представляет собой «черный ящик»). Вместе с тем известно, что в лингвистическом сознании угадчика заложены сведения о статистических свойствах языка. Речь идет, разумеется, не о численных значениях их вероятностей. На основании прошлого лингвистического опыта в сознании угадчика формируется степень убежденности по поводу того, какие лингвистические единицы встречаются чаще, а какие реже [50, стр. 260—261; 56]. Хотя «ранжирование» лингвистических элементов в сознании разных носителей языка может быть неодинаковым, оно должно приближаться при достаточно хорошем знании языка к некоторой оптимальной схеме. Поскольку испытуемый угадывает текст, опираясь каждый раз на ранжированный спектр лингвистических элементов, всю совокупность его предсказаний можно рассматривать как некоторую систему лингвистических реакций, в которой более или менее полно отражаются статистические процессы образования текста, за которыми стоят вероятностные свойства языка [ср. 57, стр. 4].

4. Если угадывание осуществляется по оптимальной схеме (идеальное угадывание), его эффективность может *быть использована в качестве меры организованности* лингвистического кода (языка) порождающего данный текст.

При проведении эксперимента используются разные приемы. Сущность всех этих приемов заключается в следующем. Берется текст, полностью или частично неизвестный для испытуемых. Испытуемые должны восстановить неизвестную им часть, последовательно отгадывая (сами или с помощью экспериментатора) буквы текста. Пробел между словами, условно обозначаемый знаком △, а также приравниваемые ему апостроф и черточка считаются конечной буквой слова.

В настоящей работе использованы данные, полученные в результате использования двух типов экспериментов. Эксперименты первого типа предусматривали угадывание текста одним испытуемым, эксперименты второго типа опирались на коллективное угадывание.

Лингвистической и информационной интерпретации подвергаются в основном статистические данные, которые были добыты с помощью двух экспериментов первого типа. Первый прием мы будем условно называть угадыванием по полной программе. Второй — угадыванием по сокращенной программе. Поскольку остальные приемы имеют контрольно-вспомогательное значение (их описание см.

[5] Текст, разумеется, можно угадывать также и по фонемам, слогам, морфемам, словам, словосочетаниям, предложениям и т. д. Предпочтительное использование буквенного алфавита объясняется его универсальностью, привычностью и понятностью для каждого грамотного носителя языка, чего нельзя сказать о фонемном, слоговом или морфемном алфавитах. В частности для большинства языков неизвестны однозначные процедуры выделения в потоке речи таких дискретных единиц, как фонемы и слоги.

Что же касается алфавитов слов, словосочетаний и предложений, то число образующих их элементарных единиц практически может быть увеличено до бесконечности. Это крайне затрудняет проведение самого эксперимента и до предела усложняет процедуру обработки результатов [ср. 10, стр. 103; 44, стр. 49—53; 60; 61, стр. 805; 63, стр. 267—272; 68а; 73, стр. 157—170; 74, стр. 39—44; 82].

в § 17), рассмотрим подробнее первые два, обращая при этом особое внимание на их математическое обоснование, а также на методику обработки получаемых результатов.

§ 3. СОКРАЩЕННАЯ ПРОГРАММА УГАДЫВАНИЯ

Угадывание по сокращенной программе заключается в том, что испытуемый называет букву, которая, по его мнению, с наибольшей вероятностью стоит на n-ом шаге текста ($n-1$ букв текста уже известны испытуемому). Если предсказание оказалось правильным, экспериментатор сообщает об этом испытуемому, и последний переходит к угадыванию следующей буквы. Если угадывание было неправильным, испытуемому называется правильная буква, и он так же, как и в первом случае, переходит к угадыванию следующей буквы. В каждом случае в протоколе указывается, правильно или неправильно была угадана буква на данном шаге текста (о различении достоверных и недостоверных продолжений см. ниже).

Рассматривая результаты этого эксперимента в плане кодирования и передачи сообщения, К. Шеннон предложил следующую интерпретацию системы связи, использующей этот вид угадывания.

Имеется исходный текст и его сокращенный вариант. Последний составлен таким образом, что угаданные с первой попытки буквы отмечены черточкой, а в случае неправильного угадывания выписаны буквы исходного текста. Сокращенный текст хотя и передает ту же семантическую информацию, что и исходный, но менее избыточен, чем этот последний (см. табл. 1).

Будем рассматривать сокращенный текст как закодированную и одновременно компрессированную форму исходного текста. Предположим также, что наш информант (им может быть не только человек, но и машина) предсказывает буквы наилучшим образом, т. е. в соответствии с истинными вероятностями их появления в тексте. Предположим, наконец, что существует двойник нашего информанта, который пользуется той же стратегией угадывания. При этих условиях можно построить систему связи, преобразовывающую на входе исходного текста в сокращенный, а затем на выходе декодирующую этот последний. Блок-схема этой системы дана на рис. 1.

Таблица 1

Исходный русский текст и текст, компрессированный по сокращенной программе угадывания

Шаги текста (n)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Исходный текст	*м*	*н*	*о*	*г*	*и*	*е*	△	*й*	*е*	*л*	*е*
Сокращенный текст	*м*	*н*	—	—	—	—	-	*о*	*е*	*л*	*е*
Шаги текста (n)	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*
Исходный текст	*г*	*а*	*т*	*ы*	△	*а*	*л*	*и*	*а*	*т*	*с*
Сокращенный текст	—	—	—	—	—	*а*	*л*	—	—	—	—
Шаги текста (n)	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*	*31*	*32*	*33*
Исходный текст	*к*	*и*	*х*	△	*и*	△	*а*	*ф*	*р*	*и*	*к*
Сокращенный текст	—	—	—	—	*и*	—	—	—	—	—	—
Шаги текста (n)	*34*	*35*	*36*	*37*	*38*	*39*	*40*	*41*	*42*	*43*	*44*
Исходный текст	*а*	*н*	*с*	*к*	*и*	*х*	△	*с*	*т*	*р*	*а*
Сокращенный текст	—	—	—	—	—	—	—	—	—	—	—
Шаги текста (n)	*45*	*46*	*47*	*48*	*49*	*50*	*51*	*52*	*53*	*54*	*55*
Исходный текст	*н*	△	*с*	*о*	*г*	*л*	*а*	*с*	*н*	*ы*	Δ
Сокращенный текст	—	—	*с*	*о*	*г*	—	—	—	—	—	—

Система работает следующим образом. На входное сравнивающее устройство подается n-ая буква исходного текста. Сюда же с входного предсказывающего устройства, которое работает, опираясь на знание $(n-1)$ буквенной цепочки, подается предсказание n-ой буквы (правильное или неправильное). Результат предсказания, будучи записанным так, как это было указано выше, образует n-ю букву сокращенного текста, которая передается дальше в канал связи. Аналогичным путем обрабатываются и следующие буквы исходного текста.

Выходное сравнивающее устройство, приняв символ сокращенного текста, запускает выходное предсказывающее устройство. Поскольку это последнее является двойником входного предсказателя, оно должно с первой же попытки восстановить отсутствующие буквы сокращенного текста — буквы, которые были правильно предсказаны первым предсказателем. Те же буквы, которые были неправильно угаданы на входе, обозначены в сокра-

щенном тексте. Восстановление исходного текста из сокращенного, которое мы будем называть обратимостью текста, указывает на отсутствие потерь смысловой информации в описанной системе связи [52, стр. 16—17].

§ 4. ПОЛНАЯ ПРОГРАММА УГАДЫВАНИЯ

Проведение полной программы угадывания также дает сокращенный текст, который мы будем называть цифровым текстом. Отличие полной программы от сокращенной заключается в том, что после неудачной попытки угадывания испытуемому не сообщается правильная буква, а он продолжает попытки предсказания вплоть до получения правильного результата. В протоколе каждый раз фиксируется число попыток, понадобившихся для того, чтобы определить, какая буква стоит на данном шаге текста. При проведении как полной, так и сокращенной программ угадывания особо выделяются достоверные продолжения. Достоверными продолжениями считаются такие, стоящие на n-ом шаге текста буквы, появление которых полностью предопределено с точки зрения современных орфографических и стилистических норм литературного языка предшествующими $n-1$ буквами текста (ср. конечное *о* и пробел в русском слове *которого* △ или буквы *e*, *r* в английском слове *other*△). Достоверными продолжениями мы будем также считать и такие буквы, появление которых на данном шаге текста имеет очень большую вероятность (т. е. при $p > 0.90$, см. об этом в § 16). В качестве примера можно указать на появление пробела после русского букворяда *мать*. Продолжение *е* (ср. французское имя собственное *Матье* или фамилию *Матьез*) вне специфических контекстов имеет очень малую вероятность.

Протокол угадывания по полной программе приводившегося выше отрывка показан в таблице 2. Здесь в строке «цифровой текст» указывается число попыток, понадобившихся для правильного отгадывания соответствующей буквы исходного текста.

Достоверные продолжения отмечаются с помощью цифры '0' [25, стр. 117].

Прежде чем анализировать результаты угадывания, рассмотрим некоторые свойства угадывания и порождаемого им сокращенного текста.

Таблица 2

Исходный текст и сокращенный (цифровой) текст, полученный в результате проведения полной программы угадывания

Шаги текста (n)	1	2	3	4	5	6	7	8	9	10	11
Исходный текст	*м*	*н*	*о*	*г*	*и*	*е*	△	*д*	*е*	*л*	*е*
Цифровой текст	8	3	1	1	1	1	0	5	3	2	2
Шаги текста (n)	12	13	14	14	16	17	18	19	20	21	22
Исходный текст	*г*	*а*	*т*	*ы*	△	*а*	*з*	*и*	*и*	*т*	*с*
Цифровой текст	1	1	1	0	0	4	7	1	0	0	0
Шаги текста (n)	23	24	25	26	27	28	29	30	31	32	33
Исходный текст	*к*	*и*	*х*	△	*и*	△	*а*	*ф*	*р*	*и*	*к*
Цифровой текст	0	0	0	0	4	1	1	1	0	0	0
Шаги текста (n)	34	35	36	37	38	39	40	41	42	43	44
Исходный текст	*а*	*н*	*с*	*к*	*и*	*х*	△	*с*	*т*	*р*	*а*
Цифровой текст	0	0	0	0	0	0	0	1	0	0	0
Шаги текста (n)	45	46	47	48	49	50	51	52	53	54	55
Исходный текст	*н*	△	*с*	*о*	*г*	*л*	*а*	*с*	*н*	*ы*	△
Цифровой текст	0	0	4	2	2	0	0	1	1	1	0

§ 5. СОКРАЩЕННЫЙ ТЕКСТ

Если рассматривать сокращенный текст как закодированную форму буквенного текста, то весь процесс кодирования можно представить следующим образом.

Имеем текст, написанный на языке L, который использует алфавит, состоящий из S букв ($A, B, C \ldots Z$, пробел). Этот текст преобразуется в цифровой текст, который, в свою очередь, написан на языке L_1, использующем алфавит, символами которого являются натуральные числа $1, 2, \ldots, k, \ldots, S$, а также число нуль. Символы этого алфавита, кроме нуля, упорядочены по убыванию их вероятностей. Это значит, что символ '1' появляется в позиции l_n i-того текста всякий раз, когда за цепочкой b_i^{n-1} следует такая буква (обозначим ее j_1), для которой $p(j_1|b_i^{n-1})$ *является максимальной. Соответственно* символ '2' ставится в позиции l_n тогда, когда за b_i^{n-1} идет буква (j_2), для которой $p(j_2|b_i^{n-1})$ является второй по величине и т. д. Достоверные продолжения (т. е. случаи, когда $p(j|b_i^{n-1}) \approx 1$, см. § 16) отмечаются символом '0'.

Что же касается сокращенной программы, то она дает текст, который записывается в алфавите, состоящем только

из трех символов: 0 — достоверное продолжение, 1 — правильное угадывание, 2 — неправильное угадывание.

Цифровой текст, так же как и сокращенный текст первого типа, может рассматриваться в качестве закодированной формы исходного текста и может быть передан по каналу с двумя идеальными предсказывающими устройствами. Первый предсказатель будет угадывать неизвестную букву в порядке убывания условных вероятностей букв, употребление которых допустимо на данном шаге текста. Предсказатель-двойник, получив цифру, соответствующую числу попыток, понадобившихся первому предсказателю, и используя распределение условных вероятностей букв, безошибочно восстановит букву исходного текста. Полная обратимость цифрового текста снова говорит об отсутствии потерь смысловой информации в нашей системе связи.

Преобразование исходного текста в сокращенный упрощает статистическую структуру текста. Это упрощение, связанное, как в этом мы сможем убедиться ниже, с ослаблением взаимозависимостей между символами, позволяет предположить, что вероятности символов цифрового текста (эти вероятности мы будем обозначать символом q) перестают зависеть или слабо зависят от i (т. е. от вида предшествующей буквенной цепочки b^{n-1}), а зависят от k и n. Если это предположение окажется справедливым, то энтропию сокращенного текста, полученного в ходе идеального угадывания, можно будет оценить с помощью одномерного распределения его символов (см. § 8). С точки зрения объема счетной работы такая задача может быть легко решена. Правда, здесь сразу же возникает новая проблема, состоящая в том, чтобы соотнести энтропию сокращенного текста с энтропией исходного текста. Если и эта задача будет выполнена, то мы получим возможность i-связную цепь Маркова языкового текста оценить с помощью нуль-связной цепи сокращенного текста [15, стр. 162].

При исследовании как отдельных слов, так и связных текстов обычно используются выборки в 100 цепочек (подробнее см. в § 15).

Результаты угадывания связных текстов и отдельных слов обобщаются в матрицы. Часть таких матриц, суммирующих результаты угадывания 100 стобуквенных русских текстов *по полной* и *сокращенной программам*,

Таблица 3

Результаты угадывания ста русских текстов по полной программе

Относительные частоты предсказания букв с определенного числа попыток на каждом шаге текста

Число попыток (k)	Шаг текста (n)										
	2	3	4	5	6	7	8	9	10	...	100
1	0.43	0.41	0.46	0.36	0.40	0.42	0.42	0.42	0.39	...	0.40
2	0.13	0.12	0.17	0.17	0.18	0.18	0.15	0.15	0.16	...	0.06
3	0.07	0.03	0.05	0.08	0.07	0.08	0.07	0.07	0.05	...	0.03
4	0.07	0.07	0.06	0.06	0.05	0.02	0.02	0.02	0.03	...	0.01
5	0.07	0.05	0.04	0.03	0.03	0.04	0.05	0.01	0.01	...	0.01
6	0.05	0.05	0.05	0.04	0.06	0.03	0.01	0.01	0.02	...	—
7	0.03	0.03	—	0.02	0.01	0.02	0.01	0.01	—	...	0.01
8	0.04	0.02	0.01	0.03	0.02	—	0.01	0.02	0.02	...	—
9	—	0.04	—	—	0.02	0.04	0.01	0.01	0.02	...	0.01
10	0.04	0.03	—	—	—	—	—	0.01	0.02	...	0.01
11	—	0.01	0.02	0.01	—	—	0.02	—	—	...	0.01
12	0.03	—	0.03	0.01	0.01	0.01	0.01	—	0.01	...	·
13	0.02	0.01	0.01	·	—	0.01	—	0.01	—	...	·
14	·	0.01	0.01	0.01	—	—	—	0.01	—	...	—
15	0.01	—	0.01	—	0.01	—	—	—	—	...	—
16	—	0.01	—	0.02	0.01	—	0.01	0.01	—	...	—
17	—	0.01	—	—	0.01	—	0.01	0.01	—	...	—
18	0.01	0.01	—	·	·	—	—	0.01	—	...	—
19	—	0.01	·	0.01	·	—	—	0.01	—	...	—
20	—	0.01	0.02	—	0.01	0.01	0.01	—	—	...	0.01
21	—	0.02	0.01	·	0.01	·	0.01	—	—	...	·
22	—	0.01	—	—	—	—	0.01	—	—	...	·
23	—	0.01	—	—	—	—	—	—	—	...	0.01
24	—	—	—	—	—	—	0.01	0.01	0.01	...	—
25	—	—	—	—	—	—	—	0.01	—	...	—
26	—	—	—	—	—	—	—	—	—	...	0.01
27	—	—	—	—	—	—	—	—	—	...	—
28	—	—	—	—	—	—	—	—	—	...	—
29	—	—	—	—	—	—	—	—	—	...	—
30	—	0.01	—	—	—	—	—	—	—	...	—
31	—	—	—	—	—	—	—	—	—	...	—
32	—	—	—	—	—	—	—	—	—	...	—
0	·	0.02	0.05	0.15	0.10	0.14	0.15	0.18	0.26	...	0.42
$\sum q_k^n$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	...	1.00
Дв. сл. H_n	2.85	3.16	2.45	2.28	2.54	2.12	2.22	2.05	1.65	...	1.02
Дв. сл. H'_n	1.97	2.18	1.63	1.54	1.60	1.33	1.44	1.34	1.04	...	0.63
Дв. сл. H''_n	1.94	2.11	1.55	1.52	1.56	1.28	1.34	1.30	1.02	...	0.58

Оценки энтропии цифрового текста

Примечание. Значения H_n и H''_n несколько изменились по сравнению с аналогичными величинами, приведенными в коллективной работе [25, стр. 127—128] в связи с тем, что уже после выхода этой работы была заново пересмотрена и в отдельных случаях изменена оценка достоверности некоторых продолжений в экспериментальных текстах.

приведены в таблицах 3 и 4. Каждый столбец в таблицах указывает на определенный шаг текста (n). Номер строки k соответствует числу попыток при угадывании. Иными словами, номер строки соответствует символу цифрового текста (ср. выше). Для полной программы число строк равно числу букв алфавита исходного языка (для русского языка $S=32$) плюс одна строка для достоверных продолжений. Для сокращенной программы задается три строки соответственно трем символам сокращенного текста (см. выше).

Число, стоящее на пересечении n-ого столбца и k-той строки, показывает условную вероятность того, что буква будет угадана с k-той попытки. Эта условная вероятность получена как частное от деления общего числа угадываний с k попыток на общее количество отрывков.

В нижних строках таблицы 3 указываются верхние с учетом достоверных продолжений оценки энтропии (H_n), нижние экспериментальные оценки (H_n'), а также нижние оценки, полученные путем приближения к идеальному угадыванию (H_n) [ср. 15, стр. 169; 25, стр. 127—128; 77, стр. 677].

Таблица 4

Результаты угадывания ста русских текстов по сокращенной программе

	Эмпирические условные вероятности										
k	n										
	2	3	4	5	6	7	8	9	10	...	100
0	—	0.02	0.05	0.15	0.10	0.14	0.15	0.18	0.26	...	0.42
1	0.43	0.41	0.46	0.36	0.40	0.42	0.42	0.42	0.39	...	0.40
2	0.57	0.57	0.49	0.49	0.50	0.44	0.43	0.40	0.35	...	0.18
Σq_k^n	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	...	1.00
H' (дв. ед.)	2.70	2.67	2.42	2.21	2.39	2.18	2.14	2.02	1.79	...	1.06
	Верхняя оценка энтропии сокращенного текста										

Данные по первому шагу текста не приводятся, поскольку его энтропия получена во всех выборках из распределения частот первых букв слова.

Сходным образом строится и таблица 4.

§ 6. ИДЕАЛЬНОЕ ПРЕДСКАЗАНИЕ

Для того чтобы оценить энтропию цифрового текста и соотнести ее с энтропией исходного текста, необходимо рассмотреть закономерности идеального предсказания букв на n-ом шаге текста в некоторой репрезентативной выборке отрывков при условии, что $n-1$ предшествующих букв всех отрывков известны предсказателю.

Поскольку идеальное предсказывающее устройство угадывает буквы на каждом шаге текста и после каждой цепочки в порядке убывания их условных вероятностей, постольку вероятность появления угадываний с первой попытки на n-ом шаге матрицы будет равно

$$q_1^n = \sum_{b_i^{n-1}} p(b_i^{n-1})\, p(j_{i,1} | b_i^{n-1}).$$

Величину q_1^n можно рассматривать и как суммарную условную вероятность появления символа '1' языка на n-ом шаге цифрового текста. Аналогичным образом задаются суммарные частоты символов '2', '3', ..., k, ..., S, обозначаемые как $q_2^n, q_3^n, \ldots, q_k^n, \ldots, q_S^n$. При этом j выбирается так, что оно дает второе, третье, ..., k-тое, ..., S-тое по величине значения p. В общем случае имеем

$$q_k^n = \sum_{b_i^{n-1}} p(b_i^{n-1})\, p(j_{i,k} | b_i^{n-1}). \qquad (7)$$

Поскольку после каждой b_i^{n-1} буквы исходного алфавита заново упорядочиваются так, что

$$p(j_{i,1} | b_i^n) \geqslant p(j_{i,2} | b_i^n) \geqslant p(j_{i,S} | b_i^n),$$

постольку каждому символу цифрового алфавита L_1 соответствует несколько букв алфавита L. В результате этого усреднения статистические связи между символами цифрового текста оказываются более слабыми, чем связи между буквами исходного текста.

При осуществлении идеального угадывания букв на следующем — $(n+1)$-ом — шаге текста получаем новое распределение вероятностей цифровых символов: q_1^{n+1}, q_2^{n+1}, ..., q_S^{n+1}. Угадывание опирается здесь на уже известную цепочку из n символов, которая, как подчеркивает Шеннон, дает больше знаний, чем цепочка в $n-1$ символов. Можно ожидать, что условные вероятности наиболее частых, а также достоверных продолжений будут возрастать, а условные вероятности редких предложений будут уменьшаться.[6] Таким образом, предсказание $(n+1)$-й буквы облегчается по сравнению с угадыванием n-ой буквы. В связи с этим предлагается следующее неравенство:

$$\sum_{k=1}^{g} q_k^{n+1} \geqslant \sum_{k=1}^{g} q_k^n, \quad g = 1, 2 \ldots . \tag{8}$$

Смысл этого неравенства состоит в том, что вероятность правильного предсказания буквы за первые g попыток, при условии, что известны предшествующие n букв, больше или равна той же вероятности, когда известны $n-1$ букв для всех g. Действительно, для любой $(n-1)$ цепочки $i_2, \ldots, i_n$ и любой буквы j в силу формулы полной вероятности имеем

$$p(i_2, i_3, \ldots, i_n, j) = \sum_{i_1=1}^{S} p(i_1, i_2, \ldots, i_n, j). \tag{9}$$

Фиксируем произвольное g.

По определению частот q_k^n неравенство (7) запишется

$$\sum_{k=1}^{g} \sum_{i_1, \ldots, i_n} p(i_1, \ldots, i_n, j_k) \geqslant \sum_{k=1}^{g} \sum_{i_2, \ldots, i_n} p(i_2, \ldots, i_n, j_k).$$

j_k — символ, который дает k-тую по величине вероятность следовать за n-ой или $(n-1)$-й цепочкой. Но

$$p(i_2, \ldots, i_n, j_k) = \sum_{i_1=1}^{S} p(i_1, i_2, \ldots, i_n, j_k). \tag{10}$$

[6] Эти соображения К. Шеннона специально рассмотрены и развиты в дипломной работе Б. А. Грошевой (1966 г., математико-механический факультет ЛГУ).

Теперь фиксируем n-грамму $i_1, i_2, \ldots, i_n$ и рассмотрим две суммы:

$$\sum_{k=1}^{g} p(i_1, \ldots, i_n, j_k)$$

и

$$\sum_{k=1}^{g} p(i_2, \ldots, i_n, j_k).$$

Первая из них — сумма g наибольших чисел из совокупности $\{p(i_1, i_2, \ldots, i_n, j)\}$ (j пробегает весь алфавит), а вторая — тоже сумма g чисел из этой совокупности, но не обязательно наибольших.

Следовательно

$$\sum_{k=1}^{g} p(i_1, \ldots, i_n, j_k) \geqslant \sum_{k=1}^{g} p(i_2, \ldots, i_n, j_k).$$

Суммируя эти неравенства по всем i и учитывая (10), получим:

$$\sum_{k=1}^{g} \sum_{i_1, \ldots, i_n} p(i_1, \ldots, i_n, j_k) \geqslant \sum_{k=1}^{g} \sum_{i_2, \ldots, i_n} p(i_2, \ldots, i_n, j_k),$$

что и нужно было доказать.

Поскольку по (7) частные суммы

$$O_g^n = \sum_{k=1}^{g} q_k^n$$

меньше единицы и монотонно не убывают при увеличении n, то при $n \to \infty$ они стремятся к пределу.

Положив $g = 1$, имеем

$$O_1^n = q_1^n,$$

т. е. q_1^n при $n \to \infty$ стремится к некоторому пределу, который мы обозначим q_1^∞. Положив $g = 2$, получим $O_2^n = q_1^n + q_2^n$, т. е. $q_2^n = O_2^n - O_1^n$, но последовательности $\{O_2^n\}$ и $\{O_1^n\}$ сходятся при $n \to \infty$ к некоторым конечным пределам, поэтому и $q_2^n \xrightarrow[n\to\infty]{} q_2^\infty$ и т. д.

Таким образом, все последовательности $\{q_k^n\}$ сходятся при $n \to \infty$ к $q_k^\infty (k = 1, 2, \ldots, S)$.

Эти пределы можно рассматривать как относительные частоты первого, второго и т. д. угадывания при условии, что известна бесконечная предыстория текста [77, стр. 681]. Из всего сказанного следует, что на n-ом шаге усредненного текста распределение значений q_k^n и сумм $\sum_{k=1}^{g} q_k^n$ всё меньше зависит от b_i^{n-1} по сравнению с распределениями вероятностей букв исходного текста и суммы этих вероятностей [ср. 15, стр. 164—165; 27, стр. 36]. Что же касается величины q_k^∞, то она вообще перестает зависеть от b_i^∞.

В тех случаях, когда условные вероятности наиболее вероятных продолжений на $(n+1)$-ом шаге текста оказываются меньше, чем соответствующие вероятности на n-ом шаге, то неравенство (7) превращается в неравенства противоположного смысла. Эта ситуация, не предусмотренная Шенноном, характерна для естественного текста, имеющего квантовую («зернистую») структуру. Кванты текста, в первую очередь слова (отчасти морфемы и устойчивые словосочетания), дают значительное увеличение условных вероятностей наиболее вероятных продолжений на конечных буквах, этот рост особенно ощутим на пробелах. Поэтому последовательность величин

$$\sum_{k=1}^{g} q_k^1,\ \sum_{k=1}^{g} q_k^2,\ \ldots,\ \sum_{k=1}^{g} q_k^n,\ \sum_{k=1}^{g} q_k^{n+1}$$

в выборке реальных текстов не образует монотонно возрастающий ряд, а будет характеризоваться перепадами. Эти перепады будут обнаруживаться между теми шагами выборки, на которых стоит большое число пробелов между словами и следующим шагом, на который соответственно падает большое число начал слов [25, стр. 121—123; 27, стр. 34—35; 35, стр. 116—117].

Однако кванты текста (слова, а также морфемы и словосочетания) неравновероятны и обладают разной длиной. Поэтому при достаточно больших выборках отрывков можно ожидать, что пробелы окажутся разнесенными по разным буквенным позициям и число перепадов будет сравнительно невелико. Вероятно, они будут ощутимы лишь в начале суммарного текста, примерно вплоть до пятнадцатой буквенной позиции. Таким образом, утверждение Шеннона о том, что неравенство (7) выполняется на всем протяжении цифрового текста, может быть использовано в качестве рабочей гипотезы при выведении формулы, описывающей распределение информации в усредненном тексте языка.

§ 7. ПРЕОБРАЗОВАНИЕ ЦИФРОВОГО ТЕКСТА В ИСХОДНЫЙ

Если рассматривать идеальное предсказывание как преобразователь, переводящий буквенный текст в последовательность чисел от 1 до S, то нетрудно заметить, что входной сигнал n-ого шага текста (определенная буква исходного текста) может быть мгновенно восстановлен путем анализа выходного цифрового символа и предшествующей цепочки из $n-1$ букв.

Представим себе таблицу, в которой по вертикали расположены все b_i^n, а по горизонтали даны цифровые символы от единицы до S. В каждой клетке, образованной пересечением столбца и строки, указана вероятность, которая одновременно соответствует и данному цифровому символу (т. е. выходному сигналу, выдаваемому входным сравнивающим устройством, см. рис. 1) и некоторой букве (входному сигналу) j_k, являющейся последним символом цепочки b_i^n, j_k. При этом следует помнить, что вероятности расположены слева направо в порядке убывания их величин. Путем соотнесения через величину вероятности цифрового символа с цепочкой b_i^n, j_k получаем букву j_k исходного текста. Задача эта имеет однозначное решение лишь в том случае, если в каждой строке нет одинаковых чисел (вероятностей). Иначе при декодировании цифрового текста возникает неопределенность в выборе между несколькими возможностями в восстановлении входной буквы [77, стр. 682].

Теперь рассмотрим соотношение вероятности цифровых символов и букв на $(n+1)$-ом шаге текста. Общая вероятность g наиболее вероятных букв на выходе (обозначим ее как $\sum_{k=1}^{g} r_k$) представляет собой просуммированную по всем строкам сумму из g наиболее вероятных букв в каждой строке. Эта сумма не может быть больше

суммы наибольших вероятностей из каждой строки. Иными словами,

$$\sum_{k=1}^{S} q_k^n \geqslant \sum_{k=1}^{S} r_k^n,^7 \tag{11}$$

поскольку оба ряда являются сходящимися:

$$\sum_{k=1}^{S} q_k^n = \sum_{k=1}^{S} r_k^n = 1,$$

их остатки находятся в следующем соотношении:

$$\sum_{k=S-g}^{S} q_k^n \leqslant \sum_{k=S-g}^{S} r_k^n. \tag{12}$$

Нетрудно заметить, что в выражении (11) множество чисел q_k^n мажорирует множество чисел r_k^n (в неравенстве (12) наоборот множество r_k^n выступает мажорантой множества q_k^n). Исходя из свойств мажорирования, все числа r_k^n в неравенстве (11) могут быть получены из чисел q_k^n «*с помощью конечного числа „переносов“*. Под переносом понимается передача вероятностей от больших q к меньшим, подобно тому как тепло течет от более нагретых к менее нагретым телам, а не наоборот» [77, стр. 682]. Кроме того, вероятности r_k^n могут быть получены из мажорирующих их q_k^n путем осреднения вида

$$r_k^n = \sum a_{kl} q_l^n,$$

где все $a_{kl} \geqslant 0$ и $\sum\limits_k a_{kl} = \sum\limits_l a_{kl}$.

§ 8. ВЕРХНЯЯ ОЦЕНКА ЭНТРОПИИ

В ходе преобразования буквенного текста в цифровой происходит некоторое осреднение вероятностей — ср. выражение (7), сопровождающееся ослаблением связей между символами текста. Всякое осреднение вероятностей ведет к увеличению энтропии [76, стр. 262—263]. Поэтому, вслед за Шенноном, можно предположить, что максимальная возможная энтропия n-ого шага цифрового текста

$$\tilde{H}_n^* = -\sum_{k=1}^{S} q_k^n \log q_k^n \tag{13}$$

может быть использована в качестве верхней оценки истинного значения энтропии на n-ом шаге буквенного текста [77, стр. 683]. Иными словами, предполагается, что

$$H_n \leqslant \tilde{H}_n^*. \tag{14}$$

Для доказательства этого предположения используем, вслед за А. П. Савчук [33, стр. 155—156], общее неравенство Иенсена:

$$p_1 f(x_1) + p_2 f(x_2) + \ldots + p_m f(x_m) \leqslant f(p_1 x_1 + p_2 x_2 + \ldots + p_m x_m),$$

где $y = f(x)$ — выпуклая в открытом промежутке от a до b функция, а $x_1, x_2, \ldots x_m$ — какие-то m значений аргумента, взятые в рассматриваемом промежутке, но не все равные между собой, $p_1, p_2, \ldots p_m$ — m положительных чисел, сумма которых равна единице.

Если принять $y = -x \log x$, то получим

$$-p_1 x_1 \log x_1 - p_2 x_2 \log x_2 - \ldots - p_m x_m \log x_m \leqslant$$
$$\leqslant -(p_1 x_1 + p_2 x_2 + \ldots + p_m x_m) \log (p_1 x_1 + p_2 x_2 + \ldots$$
$$\ldots + p_m x_m) \text{ [ср. 39, стр. 289—292],}$$

или

$$-\sum_{i=1}^{m} p_i x_i \log x_i \leqslant -\left(\sum_{i=1}^{m} p_i x_i\right) \log \sum_{i=1}^{m} p_i x_i.$$

Неравенство переходит в равенство тогда и только тогда, когда все x_i равны между собой.

Приняв $p_i = p(b_i^{n-1})$, а $x_i = p(j_{i,k}|b_i^{n-1})$, получим для k (т. е. для некоторого номера в упорядочении $j_{i,1}, j_{i,2}, \ldots$ $\ldots, j_{i,k}, \ldots, j_{i,S}$) следующее неравенство:

$$-\sum_{b_i^{n-1}} p(b_i^{n-1})\, p(j_{i,k}|b_i^{n-1}) \log p(j_{i,k}|b_i^{n-1}) \leqslant$$

$$\leqslant -\left[\sum_{b_i^{n-1}} p(b_i^{n-1})\, p(j_{i,k}|b_i^{n-1})\right] \log \sum_{b_i^{n-1}} p(b_i^{n-1})\, p(j_{i,k}|b_i^{n-1}) =$$

$$= -q_k^n \log q_k^n. \tag{15}$$

---

[7] Следует иметь в виду, что в тех случаях, когда $q_k^{n+1} = p(j_k)$ при $k = 1, 2, \ldots, S$, т. е.

$$\sum_k p(i_k^n, j_k) = \sum_{k,\,k} p(i_k^n, j_k),$$

предсказание оказывается неэффективным [15, стр. 165].

Неравенство (15) превратится в равенство тогда и только тогда, когда все $p(j_{i,k}|b_i^{n-1})$, независимо от того, после какой из $(n-1)$-ых цепочек они находятся, будут равны между собой. Это значит, что вероятность каждой из k-тых букв (эти буквы могут быть разными) в упорядочении $j_{i,1}, j_{i,2}, \ldots, j_{i,k}, \ldots, j_{i,s}$ совпадут с вероятностью k-того цифрового символа (q_k^n).

Просуммировав обе части неравенства (15) по всем k, получаем

$$-\sum_{k=1}^{S}\sum_{b_i^{n-1}} p(b_i^{n-1})\, p(j_{i,k}|b_i^{n-1}) \log p(j_{i,k}|b_i^{n-1}) =$$

$$= -\sum_{b_i^{n-1}} p(b_i^{n-1}) \sum_{k=1}^{S} p(j_{i,k}|b_i^{n-1}) \log p(j_{i,k}|b_i^{n-1}) \leqslant$$

$$\leqslant -\sum_{k=1}^{S}\left[\sum_{b_i^{n-1}} p(b_i^{n-1})\, p(j_{i,k}|b_i^{n-1})\right] \log \sum_{b_i^{n-1}} p(b_i^{n-1})\, p(j_{i,k}|b_i^{n-1}) =$$

$$= -\sum_{k=1}^{S} q_k^n \log q_k^n. \qquad (16)$$

Левая часть неравенства (16) дает истинное значение энтропии, правая — ее верхнюю оценку, иными словами, предположение Шеннона о том, что

$$H_n \leqslant H_n', \qquad (14)$$

оказывается справедливым. Упорядочение экспериментальных значений q_k^n в порядке убывания q и возрастания k не влияет на значение H_n' (ср. § 12).

Неравенство (14) превращается в равенство тогда и только тогда, когда для каждого k

$$p(j_{i,k}|b_i^{n-1}) = q_k^n.$$

В этом случае $p(j_{i,k}|b_i^{n-1})$ перестают зависеть от i, а зависят только от n и k (расположение же букв в последовательностях $j_{i,1}, j_{i,2}, \ldots, j_{i,k}, \ldots, j_{i,s}$ продолжает зависеть от b_i^{n-1}).

Поскольку верхняя оценка энтропии получается из одномерного распределения вероятностей q_k^n,[8] зависящих только от n и k и не зависящих от i, постольку расчет этой оценки в отличие от вычисления H_n (см. выше, стр. 16) является вполне разрешимой задачей.

[8] Само собой разумеется, что верхнюю оценку величины H_n можно было бы получить, используя вероятности появления букв на каждом шаге текста и не принимая при этом в расчет предшествующий контекст. Но такая оценка (обозначим ее $\bar{H}_n$) была бы заметно больше величины H_n', поскольку указанные вероятности можно получить, выравнивая вероятности цифровых символов, см. § 6, а также [27, стр. 45].

§ 9. ВЕРХНЯЯ ОЦЕНКА ЭНТРОПИИ ПРИ УЧЕТЕ ДОСТОВЕРНЫХ ПРОДОЛЖЕНИЙ

Учет достоверных продолжений заметно меняет верхнюю оценку энтропии в сторону ее приближения к значению H. Это легко видно из следующих рассуждений.

Предположим, что вероятность достоверных продолжений в позиции l_n равна q_0^n. В этом случае вероятность недостоверных продолжений текстов (т. е. угадывания с одной и более попыток) равна $1 - q_0^n$. Положив, что вероятность угадать с k попыток недостоверное продолжение относительно общего числа недостоверных продолжений равна

$$q'^n_k = \frac{q_k^n}{1 - q_0^n}$$

при $k = 1, 2, 3, \ldots, S$, мы получаем верхнюю оценку в виде

$$H_n = (1 - q_0^n)(-q'^n_1 \log q'^n_1 - q'^n_2 \log q'^n_2 - \ldots - q'^n_S \log q'^n_S),$$

где выражение

$$-q'^n_1 \log q'^n_1 - q'^n_2 \log q'^n_2 - \ldots - q'^n_S \log q'^n_S$$

представляет собой верхнюю оценку энтропии, а множитель $(1 - q_0^n)$ указывает на ее вероятность. Эта оценка, записанная в вероятностях q_k^n, имеет вид

$$\bar{H}_n = (1 - q_0^n)\left[-\sum_{k=1}^{S} \frac{q_k^n}{1 - q_0^n} \log \frac{q_k^n}{1 - q_0^n}\right].$$

В итоге верхнюю оценку энтропии при учете достоверных продолжений можно записать как

$$\bar{H}_n = (1 - q_0^n)\log(1 - q_0^n) - \sum_{k=1}^{S} q_k^n \log q_k^n. \qquad (17)$$

§ 10. СООТНОШЕНИЕ ВЕРХНИХ ОЦЕНОК ЭНТРОПИИ ПРИ УЧЕТЕ И БЕЗ УЧЕТА ДОСТОВЕРНЫХ ПРЕДЛОЖЕНИЙ

Рассматривая соотношение величин $\bar{H}_n$ и $\bar{H}_n''$, перепишем распределение вероятностей $q_1^n, q_2^n, \ldots, q_S^n$ в выражении (13) в виде $Q_1^n, Q_2^n, \ldots, Q_S^n$. Это распределение будет отличаться от распределения $q_0^n, q_1^n, q_2^n, \ldots, q_S^n$ в выражении (17) лишь тем, что $Q_1^n = q_0^n + q_1^n$. Для остальных Q_k^n выполняется соотношение $Q_k^n = q_k^n (k = 2, 3, \ldots, S)$.

В итоге оценку $\bar{H}_n''$ можно записать так:

$$\bar{H}_n'' = -(q_0^n + q_1^n)\log(q_0^n + q_1^n) - q_2 \log q_2 - \ldots - q_S^n \log q_S^n,$$

чтобы сравнить $\bar{H}_n''$ и $\bar{H}_n$ нужно сопоставить члены

$$-(q_0^n + q_1^n)\log(q_0^n + q_1^n)$$

и

$$(1 - q_0^n)\log(1 - q_0^n) - q_1^n \log q_1^n. \qquad (18)$$

Для этого снова обратимся к функции $f(x) = -x\log x$ при $x > 0$ и найдем экспериментальные значения ее аргумента, при которых первая производная обращается в нуль. Имеем

$$f'(x) = (-x\log x)' = -(\log x + \log e) = 0$$

$$\log x = -\log e,\ x = \frac{1}{e}.$$

Поскольку вторая производная $f''(x) = -\frac{1}{x}\log e$ отрицательная, экспериментальное значение аргумента $x = \frac{1}{e}$ является максимумом, а функция $f(x)$ возрастает в области $0 < x \leqslant \frac{1}{e}$.

Приравняв $x = q_0^n + q_1^n$, имеем $f(q_0^n + q_1^n) \geqslant f(q_1^n)$, а тем более $f(q_0^n + q_1^n) \geqslant f(q_1^n) - f(1 - q_0^n)$, поскольку $f(x) \geqslant 0$ для $1 \geqslant x > 0$.

Из всего сказанного видно, что при $0 < q_0^n + q_1^n \leqslant \frac{1}{e}$ $-(q_0^n + q_1^n)\log(q_0^n + q_1^n) \geqslant -q_1^n \log q_1^n + (1 - q_0^n) \times \log(1 - q_0^n)$, т. е. $\bar{H}_n'' \geqslant \bar{H}_n$.

Ясно также, что $\bar{H}_n'' = \bar{H}_n$ для $q_0^n = 0$, а при $q_0^n + q_1^n = 1$

$$\bar{H}_n'' = \bar{H}_n = 0. \qquad (19)$$

Описанное выше неравенство является гипотетически справедливым и в том случае, когда $\frac{1}{e} < q_0^n + q_1^n < 1$.

§ 11. СООТНОШЕНИЕ ИСТИННОГО ЗНАЧЕНИЯ ЭНТРОПИИ И ВЕРХНЕЙ ОЦЕНКИ ЭНТРОПИИ ПРИ УЧЕТЕ ДОСТОВЕРНЫХ ПРОДОЛЖЕНИЙ

Рассмотрим равенство (4), учитывая достоверные продолжения. В связи с этим разобьем всю совокупность цепочек b_i^{n-1} на две группы: в первую войдут те цепочки, после которых идут достоверные продолжения, во вторую — все остальные. В связи с этим выражение (4) запишется так:

$$H_n = -\sum_{k=1}^{S} \sum_{b_i^{n-1}\,(\text{I гр.})} p(b_i^{n-1})\, p(j_{i,k} | b_i^{n-1}) \log p(j_{i,k} | b_i^{n-1}) -$$

$$-\sum_{k=1}^{S} \sum_{b_i^{n-1}\,(\text{II гр.})} p(b_i^{n-1})\, p(j_{i,k} | b_i^{n-1}) \log p(j_{i,k} | b_i^{n-1}).$$

Во второй группе сохраняется соотношение $p(j_{i,1} | b_i^{n-1}) \geqslant p(j_{i,2} | b_i^{n-1}) \geqslant \ldots \geqslant p(j_{i,S} | b_i^{n-1}) \geqslant 0$ для любой b_i^{n-1}, причем $p(j_{i,1} | b_i^{n-1}) < 1$.

В цепочках первой группы $p(j_{i,1} | b_i^{n-1}) = 1$, а $p(j_{i,k} | b_i^{n-1}) = 0$ при $k = 2, 3, \ldots, S$.

Отсюда ясно, что энтропия первой группы цепочек равна нулю. Вместе с тем отметим, исходя из (7), что

$$\sum_{b_i^{n-1}\,(\text{I гр.})} p(b_i^{n-1})\, p(j_{i,1} | b_i^{n-1}) = q_0^n = \sum_{b_i^{n-1}\,(\text{I гр.})} p(b_i^{n-1}).$$

Истинное значение энтропии должно определяться теперь исходя из следующих соображений. Вероятность появления цепочек, включающих недостоверные продолжения j_k, равна $(1 - q_0^n)$. Вероятности появления j_k-того продолжения после цепочек второй группы равна

$$\frac{p(j_{i,k} \mid b_i^{n-1})}{1 - q_0}.$$

Заменив выражения $\sum_{b_i^{n-1}\,(\text{II гр.})} p(b_i^{n-1})\, p(j_{i,k} \mid b_i^{n-1})$ на q_k^n и $p(j_{i,k} \mid b_i^{n-1})$ на P_k^n имеем

$$H_n = (1 - q_0^n)\left[-\sum_{k=1}^{S} \frac{q_k^n}{1 - q_0^n} \log \frac{P_k^n}{1 - q_0^n}\right].$$

Отсюда получаем

$$H_n = (1 - q_0^n) \log (1 - q_0^n) - \sum_{k=1}^{S} q_k^n \log P_k^n.$$

Возвратившись к первоначальным обозначениям, имеем

$$H_n = \left[1 - \sum_{b_i^{n-1}\,(\text{I гр.})} p(b_i^{n-1})\right] \log \left[1 - \sum_{b_i^{n-1}\,(\text{I гр.})} p(b_i^{n-1})\right] -$$

$$- \sum_{b_i^{n-1}\,(\text{II гр.})} p(b_i^{n-1}) \sum_{k=1}^{S} p(j_{i,k} \mid b_i^{n-2}) \log p(j_{i,k} \mid b_i^{n-2}) \leqslant$$

$$\leqslant \left[1 - \sum_{b_i^{n-1}\,(\text{I гр.})} p(b_i^{n-1})\right] \log \left[1 - \sum_{b_i^{n-1}\,(\text{I гр.})} p(b_i^{n-1})\right] -$$

$$- \sum_{k=1}^{S} \left[\sum_{b_i^{n-1}\,(\text{II гр.})} p(b_i^{n-1})\, p(j_{i,k} \mid b_i^{n-1})\right] \times$$

$$\times \log \sum_{b_i^{n-1}\,(\text{II гр.})} p(b_i^{n-1})\, p(j_{i,k} \mid b_i^{n-1}) = (1 - q_0^n) \log (1 - q_0^n) -$$

$$- \sum_{k=1}^{S} q_k^n \log q_k^n = H_n. \qquad (20)$$

Таким образом, учет достоверных продолжений приводит нас к неравенству

$$H_n \leqslant H_n, \qquad (20а)$$

аналогичному неравенству (14). При этом верхние оценки заметно улучшаются, приближаясь к истинному значению энтропии [25, стр. 129]. Неравенство (20а) превращается в равенство при тех же условиях, что и неравенство (14).

§ 12. НИЖНЯЯ ОЦЕНКА ЭНТРОПИИ

Для определения нижней оценки энтропии используются следующие рассуждения.

Как уже не раз говорилось, после каждой цепочки b_i^{n-1} идеальный угадыватель расставляет буквы алфавита на n-ом шаге текста в порядке убывания их вероятностей таким образом, что

$$p(j_1 \mid b_i^{n-1}) \geqslant p(j_2 \mid b_i^{n-1}) \geqslant \ldots \geqslant p(j_k \mid b_i^{n-1}) \geqslant$$
$$\geqslant p(j_{k+1} \mid b_i^{n-1}) \geqslant \ldots \geqslant p(j_S \mid b_i^{n-1}).$$

Это неравенство после некоторой конкретной $(n-1)$-ой цепочки будет иметь вид

$$p_1^n \geqslant p_2^n \geqslant \ldots \geqslant p_k^n \geqslant p_{k+1}^n \geqslant \ldots \geqslant p_S^n. \qquad (20б)$$

Напомним, что

$$\sum_{k=1}^{S} p_k^n = 1, \quad 0 \leqslant p_k^n \leqslant 1, \quad p_{S+1}^n = 0.$$

Каждую величину p_k^n можно представить в виде суммы

$$p_k^n = (p_k^n - p_{k+1}^n) + (p_{k+1}^n - p_{k+2}^n) + \ldots + (p_{S-1}^n - p_S^n) +$$
$$+ (p_S^n - p_{S+1}^n) = \sum_{i=k}^{S} (p_i^n - p_{i+1}^n),$$

а все p_k^n объединить в матрицу (см. таблицу 5).

Просуммировав строки матрицы, получаем

$$\sum_{k=1}^{S} (p_k^n - p_{k+1}^n) = \sum_{k=1}^{S} p_k^n = 1.$$

В каждом k-том заранее заданном столбце содержится k равновероятных членов. Поэтому энтропия бу-

Таблица 5

Матрица разностей $(p_k^n - p_{k+1}^n)$

x	$y=k$					
	1	2	...	k	...	S
1	$p_1^n = (p_1^n - p_2^n) + (p_2^n - p_3^n) + \ldots + (p_k^n - p_{k+1}^n) + \ldots + (p_S^n - p_{S+1}^n)$					
2	$p_2^n =$	$(p_2^n - p_3^n) + \ldots + (p_k^n - p_{k+1}^n) + \ldots + (p_S^n - p_{S+1}^n)$				
...	...					
k	$p_k^n =$			$(p_k^n - p_{k+1}^n) + \ldots + (p_S^n - p_{S+1}^n)$		
...	...					
S	p_S^n					$(p_S^n - p_{S+1}^n)$

дет равна здесь $\log k$. Вероятность k-того столбца в матрице $p(y=k) = k(p_k^n - p_{k+1}^n)$.

Поскольку $y=k$ каждый раз задано, средняя энтропия распределения выступает в виде средней условной энтропии

$$H(x \mid y) = \sum_{k=1}^{S} k (p_k^n - p_{k+1}^n) \log k,$$

в то время как безусловная энтропия распределения p_k^n в нашей матрице равна

$$H(x) = -\sum_{k=1}^{S} p_k^n \log p_k^n.$$

Поскольку $H(x) \geqslant H(x \mid y)$ [ср. 39, стр. 291—292; 76, стр. 262—263], то

$$-\sum_{k=1}^{S} p_k^n \log p_k^n \geqslant \sum_{k=1}^{S} k (p_k^n - p_{k+1}^n) \log k.$$

В полной записи это выглядит так:

$$-\sum_{k=1}^{S} p(j_{i,k} \mid b_i^{n-1}) \log p(j_{i,k} \mid b_i^{n-1}) \geqslant$$

$$\geqslant \sum_{k=1}^{S} k [p(j_{i,k} \mid b_i^{n-1}) - p(j_{i,s+1} \mid b_i^{n-1})] \log k. \qquad (21)$$

Распространяя эти рассуждения относительно распределений букв на n-ом шаге текста после всех b_i^{n-1}, получаем

$$H_n = -\sum_{b_i^{n-1}} p(b_i^{n-1}) \sum_{k=1}^{S} p(j_{i,k} \mid b_i^{n-1}) \log p(j_{i,k} \mid b_i^{n-1}) \geqslant$$

$$\geqslant \sum_{b_i^{n-1}} p(b_i^{n-1}) \sum_{k=1}^{S} k [p(j_{i,k} \mid b_i^{n-1}) - p(j_{i,k+1} \mid b_i^{n-1})] \log k =$$

$$= \sum_{k=1}^{S} k \log k \left[\sum_{b_i^{n-1}} p(b_i^{n-1}) p(j_{i,k} \mid b_i^{n-1}) - \right.$$

$$\left. - \sum_{b_i^{n-1}} p(b^{n-1}) p(j_{i,k+1} \mid b_i^{n-1}) \right] = \underline{H}_n.$$

Помня, что, согласно (7), каждое

$$\sum_{b_i^{n-1}} p(b_i^{n-1}) p(j_{i,k} \mid b_i^{n-1}) = q_k^n,$$

получаем

$$H_n \geqslant \sum_{k=1}^{S} k (q_k^n - q_{k+1}^n) \log k = \underline{H}_n. \qquad (22)$$

Упорядочение значений q_k^n по убыванию q и возрастанию k уменьшает значение $\underline{H}_n$.

Истинное значение энтропии тогда совпадает с нижней оценкой, когда для каждого b_i^{n-1} при условии, что $p(b_i^{n-1}) > 0$, неравенство (21) превращается в равенство. Это происходит тогда и только тогда, когда $p(j_{j,k} \mid b_i^{n-1}) = \frac{1}{m_i}$ при $k \leqslant m_i$ и $p(j_{i,k} \mid b_i^{n-1}) = 0$ при $k > m_i$, $m_i = 1, 2, \ldots, S$.

В этом случае

$$H_n = \underline{H}_n = \sum_{b_i^{n-1}} p(b_i^{n-1}) \log m_i. \qquad (23)$$

Учет достоверных продолжений не изменяет формулы (22): снова положив, что вероятность достоверных продолжений составляет q_0^n, а вероятность появления новой оценки нижней границы равняется $1 - q_0^n$, получаем

$$\underline{H}_n = (1 - q_0^n) \sum_{k=1}^{S} k \frac{(q_k^n - q_{k+1}^n)}{1 - q_0^n} \log k = \sum_{k=1}^{S} k (q_k^n - q_{k+1}^n) \log k.$$

Верхняя и нижняя оценки сходятся с истинным значением только тогда, когда для любого b_i^{n-1} вероятность

$$p(j_{i,k} \mid b_i^{n-1}) = \begin{cases} \frac{1}{m_i} & \text{при } k \leqslant m_i \\ 0 & \text{при } k > m_i, \text{ где } 1 \leqslant m_i \leqslant S. \end{cases}$$

В этом случае

$$\bar{H}_n = H_n = \underline{H}_n = \log m_i. \tag{24}$$

Если при тех же условиях учитывать достоверные продолжения, вероятность которых равна вероятности каждого из $p(j_{i,k} \mid b_i^{n-1})$, то

$$\bar{H}_n = H_n = \underline{H}_n = \frac{m}{m+1} \log m. \tag{25}$$

Само собой разумеется, что условия, при которых оказываются справедливыми равенства (23, 24, 25), в естественных языках никогда не выполняются.

§ 13. РАСЧЕТ ВЕРХНЕЙ ОЦЕНКИ ЭНТРОПИИ ПО СОКРАЩЕННОЙ ПРОГРАММЕ УГАДЫВАНИЯ

При обработке результатов сокращенной программы угадывания с учетом достоверных продолжений верхняя оценка энтропии ($\bar{H}'_n$) представляет собой сумму из двух членов. Первый член указывает на полученную испытуемым с вероятностью $(1-q_0^n)$ информацию о том, что первое его угадывание правильно или неправильно. Эта информация численно равна

$$-q_1'^n \log q_1'^n - (1-q_1'^n) \log (1-q_1'^n).$$

Второй член представляет собой информацию, получаемую испытуемым с вероятностью $(1-q_0^n-q_1^n)$ при сообщении ему экспериментатором правильной буквы в том случае, когда не было достоверного продолжения и первая попытка угадывания не удалась. Эта информация обычно меньше или равна информации, получаемой после правильного угадывания буквы текста при условии, что испытуемому были известны две предшествующие буквы (начиная с третьей буквы текста, испытуемому известно по крайней мере две буквы, предшествующие угадываемой букве). Последняя информация и равна количественно H_{III}. В результате всех этих рассуждений мы получаем

$$\bar{H}'_n = (1-q_0^n)\left[-q_1'^n \log q_1'^n - (1-q_1'^n) \log (1-q_1'^n)\right] + (1-q_0^n-q_1^n) H_{\text{III}}.$$

Заменяя вероятности $q_k'^n$ на q_k^n, мы приходим к выражению:

$$\bar{H}'_n = H_{\text{III}}(1-q_0^n-q_1^n) - q_1^n \log q_1^n + (1-q_0^n) \log (1-q_0^n) - (1-q_0^n-q_1^n) \log (1-q_0^n-q_1^n). \tag{26}$$

§ 14. СООТНОШЕНИЕ ВЕРХНИХ ОЦЕНОК ЭНТРОПИИ, ПОЛУЧЕННЫХ ПО ПОЛНОЙ И СОКРАЩЕННОЙ ПРОГРАММАМ С УЧЕТОМ ДОСТОВЕРНЫХ ПРОДОЛЖЕНИЙ

Чтобы определить, в какой степени сокращенный эксперимент может быть использован для определения верхней оценки энтропии, необходимо исследовать соотношения величин $\bar{H}'_n$ и $\bar{H}_n$, иными словами, найти условия, при которых $\bar{H}'_n > \bar{H}_n$, $\bar{H}'_n \leqslant \bar{H}_n$ и $\bar{H}'_n = \bar{H}_n$.[9] Дело в том, что в тех случаях, когда $\bar{H}'_n < \bar{H}_n$, оценку верхней границы энтропии с помощью величины $\bar{H}'_n$ проводить нельзя, поскольку может оказаться, что $\bar{H}'_n < H_n$.

Пусть выполняется неравенство

$$\bar{H}_n < \bar{H}'_n, \tag{27}$$

где

$$\bar{H}_n = (1-q_0^n) \log (1-q_0^n) - \sum_{k=1}^{S} q_k^n \log q_k^n,$$

при

$$q_k^n = \sum_{b_i^{n-1}} p(b_i^{n-1})\, p(j_{i,k} \mid b_i^{n-1})$$

и

$$p(j_{i,1} \mid b_i^{n-1}) \geqslant p(j_{i,2} \mid b_i^{n-1}) \geqslant \ldots \geqslant p(j_{i,S} \mid b_i^{n-1})$$

для любой b_i^{n-1}, откуда

$$q_1^n \geqslant q_2^n \geqslant \ldots \geqslant q_S^n.$$

---

[9] Приводимые ниже соотношения оценок энтропии были рассмотрены в дипломной работе Л. А. Гусаковой (1966 г., математико-механический факультет ЛГУ).

Чтобы оценить сверху $\bar{H}_n$, достаточно оценить сверху слагаемое $\sum_{k=1}^{S} q_k^n \log q_k^n$, для которого имеем:

$$\sum_{k=1}^{S} q_k^n \log q_k^n \leqslant \sum_{k=1}^{S} q_k^n \log q_1^n = (1 - q_0^n) \log q_1^n.$$

Отсюда

$$\bar{H}_n \geqslant (1 - q_0^n) \log (1 - q_0^n) - (1 - q_0^n) \log q_1^n.$$

Неравенство (27) принимает тогда вид

$$(1 - q_0^n) \log (1 - q_0^n) - (1 - q_0^n) \log q_1^n < (1 - q_0^n) \log (1 - q_0^n) -$$
$$- q_1^n \log q_1^n - (1 - q_0^n - q_1^n) \log (1 - q_0^n - q_1^n) + H_{\text{III}} (1 - q_0^n - q_1^n).$$

Записав выражение $(1 - q_0^n - q_1^n)$ как $\sum_{k=2}^{S} q_k^n$, получаем после преобразований

$$-\sum_{k=2}^{S} q_k^n \log q_1^n < \sum_{k=2}^{S} q_k^n H_{\text{III}} - \sum_{k=2}^{S} q_k^n \log (1 - q_0^n - q_1^n).$$

Откуда имеем

$$0 < \left(\sum_{k=2}^{S} q_k^n \right) [\log q_1^n + H_{\text{III}} - \log (1 - q_0^n - q_1^n)].$$

Пусть $\sum_{k=2}^{S} q_k^n > 0$, т. е. из рассмотрения исключим случаи, когда $q_0^n + q_1^n = 1$. Тогда

$$[\log q_1^n + H_{\text{III}} - \log (1 - q_0^n - q_1^n)] > 0$$

и

$$H_{\text{III}} - \log (1 - q_0^n - q_1^n) > - \log q_1^n. \tag{28}$$

Неравенство (28) является необходимым условием для выполнения неравенства (27) при условии, что $q_0^n + q_1^n \neq 1$.

Теперь найдем достаточное условие для выполнения неравенства (27), предварительно переписав его в следующем виде:

$$-\sum_{k=2}^{S} q_k^n \log q_k^n < H_{\text{III}} \sum_{k=2}^{S} q_k^n - \sum_{k=2}^{S} q_k^n \log \sum_{k=2}^{S} q_k^n$$

$$-\sum_{k=2}^{S} q_k^n \log q_k^n < -\sum_{k=2}^{S} q_k^n \log \left(\frac{1}{a^{H_{\text{III}}}} \sum_{k=2}^{S} q_k^n \right), \tag{29}$$

где a — основание логарифма.

Пусть $\sum_{k=2}^{S} q_k^n > 0$, тогда можно утверждать, что если выполняется неравенство

$$\frac{1}{a^{H_{\text{III}}}} \sum_{k=2}^{S} q_k^n < q_k^n, \text{ для } q_k^n \neq 0 \ (k = 2, 3, \ldots, S), \tag{30}$$

то выполняется и неравенство (29).

Действительно, логарифмируя (30) и умножая обе его части на $q_k^n \neq 0$ $(k = 2, 3, \ldots, S)$, имеем для таких q_k^n следующее неравенство:

$$q_k^n \log_a \left[\frac{1}{a^{H_{\text{III}}}} (1 - q_0^n - q_1^n) \right] < q_k^n \log_a q_k^n. \tag{31}$$

Просуммировав (31) и добавив к обеим частям нулевые слагаемые, появляющиеся в том случае, когда $q_k^n = 0$, получаем для всех $k = 2, 3, \ldots, S$:

$$\sum_{k=2}^{S} q_k^n \log_a \left[\frac{1}{a^{H_{\text{III}}}} (1 - q_0^n - q_1^n) \right] < \sum_{k=2}^{S} q_k^n \log_a q_k^n. \tag{32}$$

Умножив обе части неравенства (32) на -1, приходим к неравенству (29).

Таким образом, при условии, что $q_0^n + q_1^n \neq 1$, неравенство (30) является достаточным для выполнения неравенства (27)

Прологарифмировав (30) и умножив обе части на -1, имеем:

$$H_{\text{III}} - \log (1 - q_0^n - q_1^n) > - \log q_k^n \tag{33}$$

при $k = 2, 3, \ldots, S$ и $q_k^n \neq 0$.

Заметим, что

$$-\log q_1^n \leqslant -\log q_2^n \leqslant \ldots \leqslant -\log q_k^n \leqslant \ldots \leqslant -\log q_S^n. \tag{34}$$

Сопоставляя равенства (28) и (34), в итоге находим, что необходимым и достаточным условием выполнения

неравенства (в случае $q_0^n + q_1^n \neq 1$) является следующее соотношение:

$$H_{\text{III}} - \log(1 - q_0^n - q_1^n) > -\log q_{\min}^n,$$

где $q_{\min}^n = \min\limits_{2 \leqslant k \leqslant S} q_k^n$ при $q_k^n \neq 0$.

* * *

Обращаясь к исследованию неравенства

$$H_n > H_n', \tag{35}$$

преобразуем его согласно (29) так:

$$-\sum_{k=2}^{S} q_k^n \log_a q_k^n > -\sum_{k=2}^{S} q_k^n \log_a \left(\frac{1}{a^{H_{\text{III}}}} \sum_{k=2}^{S} q_k^n \right). \tag{36}$$

При этом утверждаем, что если выполняется неравенство

$$\frac{1}{a^{H_{\text{III}}}} \sum_{k=2}^{S} q_k^n > q_k^n \tag{37}$$

для $q_k^n \neq 0$ ($k = 2, 3, \ldots, S$) и

$$\sum_{k=2}^{S} q_k^n > 0,$$

то выполнено неравенство (36), а значит и (35).

Доказательство этого неравенства аналогично доказательству неравенства (30).

Логарифмируя (37) и помня условие (34), приходим к $H_{\text{III}} - \log(1 - q_0^n - q_1^n) < -\log q_2^n$, которое является достаточным условием для выполнения неравенства (35), если $q_0^n + q_1^n \neq 1$.

* * *

Рассмотрим соотношение:

$$H_n = H_n'. \tag{38}$$

Перепишем его в виде

$$-\sum_{k=2}^{S} q_k^n \log q_k^n = -\sum_{k=2}^{S} q_k^n \log \left(\frac{1}{a^{H_{\text{III}}}} \sum_{k=2}^{S} q_k^n \right).$$

Для выполнения этого равенства достаточно, чтобы

$$\log q_k^n = \log \frac{1}{a^{H_{\text{III}}}} \sum_{k=2}^{S} q_k^n,$$

или

$$q_k^n = \frac{1 - q_0^n - q_1^n}{a^{H_{\text{III}}}}; \quad k = 2, 3, \ldots, S; \quad q_0^n + q_1^n \neq 1.$$

* * *

Рассмотрим случай, когда

$$q_0^n + q_1^n \neq 1,$$

а

$$q_2^n = q_3^n = \ldots = q_S^n = \frac{1 - q_0^n - q_1^n}{S - 1}.$$

Пусть выполняется неравенство

$$H_n < H_n',$$

тогда, исходя из (29), имеем

$$-(S-1) q_2^n \log q_2^n > -(S-1) q_2^n \log \frac{(S-1) q_2^n}{a^{H_{\text{III}}}},$$

откуда получаем:

$$H_{\text{III}} < \log(S-1). \tag{39}$$

Это неравенство будет в данном случае необходимым и достаточным условием выполнения неравенства (27).

Аналогичным образом оказывается, что неравенство $H_{\text{III}} = \log(S-1)$ будет необходимым и достаточным условием для выполнения соотношения (38).

Напоминаем, что во всех этих случаях $q_0^n + q_1^n \neq 1$, а $q_2^n = q_3^n = \ldots = q_S^n = \frac{1 - q_0^n - q_1^n}{S-1}$.

Рассмотрим случай, когда $q_0^n + q_1^n + q_2^n = 1$, а $0 < q_0^n$, q_1^n, $q_2^n < 1$.

Имеем:

$$H_n = (1 - q_0^n) \log(1 - q_0^n) - q_1^n \log q_1^n - q_2^n \log q_2^n$$

$$H_n' = (1 - q_0^n) \log(1 - q_0^n) - q_1^n \log q_1^n - q_2^n \log q_2^n + H_{\text{III}} q_2^n.$$

Пусть $H_n > H_n'$, тогда $0 > H_{\text{III}} q_2^n$, что никогда не имеет места, поскольку $q_2^n > 0$ и $H_{\text{III}} > 0$; если $H_n = H_n'$,

то $H_{III}=0$, что также никогда не имеет места. Иными словами, в рассматриваемом случае всегда имеет место неравенство $H_n'<H_n'$, при котором выполняется условие $H_{III}>0$.

Что же касается случая, когда

$$q_0^n+q_1^n=1,$$

то

$$H_n'=H_n=H_n''=0.$$

Соотношения только что рассмотренных оценок энтропии таковы, что H_n' значительно превышает H_n в тех случаях, когда $\sum_{k=2}^{S} q_k^n$ включает большое число слагаемых ненулевого значения и когда величины q_k^n ($k=2, 3, \ldots, S$) заметно отличаются друг от друга. Этим, в частности, следует объяснять большие расхождения в величинах H_n' и H_n, полученных Н. В. Петровой для публицистических текстов французского языка, ср. [69, стр. 138] и таблицу 2.

В этом случае сблизить оценки H_n' и H_n можно, уменьшив коэффициент H_{III}. Вместо H_{III} можно взять величину H_{IV} или H_V, предполагая, что информация, получаемая испытуемым с вероятностью $(1-q_0^n-q_1^n)$, при отгадывании правильной буквы после неудачного угадывания меньше или равна информации, получаемой после правильного угадывания буквы текста при условии, что испытуемому были известны три (или четыре) предшествующие буквы. Действительно, начиная с четвертой (пятой) буквы текста, испытуемому известно по крайней мере три (четыре) буквы, предшествующих угадываемой букве.

Однако уменьшение величины указанного коэффициента связано с определенными неудобствами.

Введение постоянного члена H_{III} (или H_{IV}, H_V и т. д.) ограничивает максимальное значение энтропии, которое можно получить, используя это выражение. Чтобы показать это, используем рассуждения, предложенные Н. В. Петровой. Максимального значения H_n' достигает при условии, что $q_0=0$. Выразив член $(1-q_1)$ через q_2, найдем такие значения q_1 и q_2, при которых реализуется

экстремальное значение $H_n'(q_0=0)$. Для этого, приравняв нулю частную производную

$$\frac{\partial H_n'(q_0=0)}{\partial q_2}$$

и выразив q_1 через $(1-q_2)$, получаем выражение

$$\frac{\partial H_n'(q_0=0)}{\partial q_2}=\frac{\partial}{\partial q_2}\left[H_{III}q_2-(1-q_2)\log_a(1-q_2)-q_2\log_a q_2\right]=$$
$$=H_{III}-\log_a\left(\frac{q_2}{1-q_2}\right)=0,$$

откуда мы получаем

$$q_2=\frac{a^{H_{III}}}{1+a^{H_{III}}},$$

$$q_1=\frac{1}{1+a^{H_{III}}}.$$

Максимальные значения H_n' при $q_0=0$ и $a=2$ для различных языков даны в таблице 6.

Вместе с тем оказалось, что описанные недостатки сокращенной программы угадывания не настолько серьезны, чтобы отказаться от использования этого метода. Так, например, обследование русского и французского материала показало, что пиковые значения верхних оценок энтропии при $n>4$ никогда не достигают значений $H_n'^{max}(q_0=0)$. Одновременно было выяснено, что расхождения в частных значениях H_n и H_n' не превышают в среднем 5%. Этими отклонениями можно пренебречь, поскольку они, как правило, не искажают общей картины распределения информации в тексте, морфеме, слоге (см. ниже, стр. 72). Все это дает нам право использовать упрощенную формулу [24, стр. 112] для оценки верхних границ интервала, в которых заключено истинное значение энтропии для различных участков текста и отдельно взятого слова.

Таблица 6

Максимальные значения H' (в дв. ед.)

Языки	H_{III}	H'^{max}
Русский	3.00	3.19
Английский	2.84	3.03
Французский	2.83	3.02
Румынский	2.69	2.90

§ 15. ПОДБОР ЭКСПЕРИМЕНТАЛЬНОГО МАТЕРИАЛА

Чтобы получить усредненные вероятностные и энтропийные характеристики сокращенного текста, необходимо провести эксперимент по угадыванию в достаточно репрезентивной выборке текстов. Наблюдения показывают, что уже выборка в 75 текстах дает сравнительно небольшой разброс, а дальнейшее добавление числа исследуемых отрывков или отдельных слов незначительно сглаживают статистические колебания [15, стр. 172; 25, стр. 120].[10] Поэтому при исследовании как отдельных слов, так и связных текстов обычно используются выборки, содержащие не менее 100 цепочек.[11]

При исследовании энтропии текста в русском, польском, азербайджанском, английском, французском, а также румынском и молдавском языках использовалось 100—120 связных отрывков длиной от ста до двухсот букв. В каждом тексте угадыванию подвергались буквы от 2-й до 100-й включительно. Что касается английского, польского и азербайджанского языков, для которых обрабатывались тексты длиной не менее двухсот букв, то, кроме девяноста девяти первых букв, здесь угадывались также 150-я и 200-я буквы (отрывки от 101-й до 149-й и от 151-й до 199-й букв предварительно сообщались информанту). Экспериментально тексты обычно представляют собой простые распространенные или сложные предложения средней длины.

[10] Г. П. Богуславская [8] показала, что нижний предел объема выборки для получения достоверных информационных данных относительно английского языка равен пятидесяти текстам.

[11] Исключения составляют длинные слова, а также латинские тексты. Ввиду того что число длинных слов (от 9—10 букв и выше) в каждом языке обычно не очень велико и они имеют сходную морфологическую структуру, мы ограничивались выборками, состоящими из пятидесяти слов. Трудности эксперимента по угадыванию букв мертвого языка заставили нас ограничиться здесь исследованием пяти стобуквенных текстов. Угадывание английских текстов осуществлено в рамках группы «Статистика речи» Г. П. Богуславской, французских — Н. В. Петровой, польских — А. В. Гут, азербайджанских — Х. З. Багировой, русских — автором этих строк; эксперимент по румынскому и молдавскому языкам проведен Л. А. Новак, И. В. Матковским и автором этих строк; исследование латинских текстов осуществлено автором вместе с А. В. Грошевой. В настоящее время ведутся исследования энтропии испанского, немецкого, новогреческого, корейского и узбекского языков.

Кроме того, для каждого из новых языков было специально угадано и обработано от пятисот до шестисот отдельных слов, взятых вне контекста.

Для языкознания интерес представляют не только абсолютные величины энтропии, информации и избыточности, но также относительное распределение этих величин в различных участках текста. Абсолютные величины информации и избыточности могут быть оценены путем применения полной программы угадывания. Распределение этих величин может быть показано с помощью сокращенной программы. Комбинированное использование обеих программ значительно сокращает объем эксперимента и дает два взаимоконтролируемых ряда числовых данных. По сокращенной программе исследовались английские, французские, румынские, польские, азербайджанские, латинские тексты и отдельные слова. Это исследование сопровождалось выборочным использованием полной программы, по которой угадывались 5-я, 10-я, 15-я, 20-я, 30-я, 40-я, 50-я, 75-я, 100-я, а для английского, польского и азербайджанского языков 150-я и 200-я буквы текста. Русские тексты и слова угадывались по полной программе.

В большинстве математических работ, посвященных информационным измерениям языка, исследуются тексты, извлеченные из одного и того же произведения. Например, К. Шеннон использовал отрывки из книги «Виргинец Джефферсон» Дюма Малона, ученики А. Н. Колмогорова проводили угадывание по повести «Детские годы Багрова-внука» С. Т. Аксакова. Подбирая однородные с точки зрения жанра и индивидуального стиля тексты, авторы этих исследований стремились сократить разброс выходных данных. Однако при перенесении получаемых результатов на язык в целом возникают серьезные затруднения. Дело в том, что не существует лингвистических и статистических приемов, с помощью которых можно было бы показать, что язык того или иного художественного произведения достаточно точно отражает информационные характеристики языка в целом. Между тем общеязыковые информационные характеристики могут оказаться весьма полезными при расчете памяти специализированных электронно-вычислительных машин. Эти характеристики необходимы и при составлении таких программ автоматического анализа и синтеза текста, которые

ориентированы не на какой-либо заранее намеченный отрывок, но на любой текст, написанный на данном языке или его стилевой разновидности.

Поэтому выбор экспериментального материала осуществляется из разных источников с соблюдением следующих жанрово-стилистических пропорций:

1) устно-разговорная речь	— 30%
2) беллетристика	— 30%
3) публицистика и научно-деловая речь	— 30%
4) поэзия	— 10%
всего	— 100%

При определении информационных характеристик языка в целом из общего числа исследованных отрывков отбиралось 100 текстов указанной длины, которые распределялись в только что названных соотношениях.

Информационные характеристики по каждому из первых трех жанрово-стилистических разновидностей получены путем обработки 33 текстов, принадлежащих данной разновидности. Поэтические тексты в особую выборку не выделялись, поскольку их угадывание требует особой организации эксперимента (специальный выбор испытуемого, использование словарей рифм, таблиц частот вариантов ритма для данного метра и т. п.). Более частные детали подбора экспериментального материала и организации опытов см. в следующих работах [7, стр. 10—11; 25, стр. 115—130].

§ 16. ИНФОРМАНТ И ИДЕАЛЬНОЕ ПРЕДСКАЗАНИЕ

Кибернетическое устройство, способное осуществить идеальное предсказание букв текста, пока еще не создано. Информант, даже обладающий самым высоким лингвистическим чутьем и языковой культурой, не может дать идеального в полном смысле этого слова угадывания. Результаты предсказаний зависят здесь и от внимания испытуемого, и от степени его усталости, и от его настроения [25, стр. 117; 68, стр. 141—143; ср. также 84].

Вместе с тем существуют приемы лингвистического, психологического и математического порядка, с помощью которых можно улучшить результаты угадывания и приблизить их к результатам идеального предсказания. Наиболее эффективными можно считать следующие приемы.

1. Подбор испытуемого, обладающего хорошим чутьем языка и высокой лингвистической культурой. В описываемых экспериментах участвовали люди с высшим (для английского языка — незаконченным высшим) образованием, для которых исследуемый язык является родным и которые дают лучшие результаты в пробных отборочных экспериментах.

2. Угадывание небольших объемов текста с целью предупредить притупление лингвистической реакции испытуемого.

3. Использование вспомогательного словарно-статистического аппарата.

При угадывании букв незнакомого текста информант обычно пользуется:

а) специально составленными таблицами частотности начальных букв и двухбуквенных сочетаний, а также таблицами частотности двухбуквенных и трехбуквенных сочетаний независимо от позиций этих последних в слове; таблицы частотности начальных букв используются при угадывании первой буквы слова, остальные таблицы применяются при угадывании второй и третьей букв;

б) энциклопедическими и двуязычными словарями, которые используются при угадывании четвертой и последующих букв, а также при определении достоверных продолжений. Применение статистических таблиц и словарей не только облегчает и ускоряет процесс угадывания, но в значительной степени нивелирует отклонения в результатах эксперимента у разных испытуемых. Об этом можно судить, в частности, по небольшим отклонениям контрольного эксперимента (ср. ниже).

4. Вероятностно-лингвистическая коррекция протокола.

Первый вид коррекции, осуществляемый экспериментатором совместно с испытуемым, имеет целью устранить «лишние попытки» при угадывании отдельных букв. Смысл устранения лишних попыток можно проиллюстрировать на следующем русском примере. В тексте *он не литера-тор. . .* (газ. «Известия» от 20-го ноября 1960 г.) следующая за вторым *т* буква была угадана со второй попытки,

что и было указано в протоколе. Последующая проверка по словарям показала, что в данном случае имеет место система из двух выборов: *литерату(-ра, -рный* и т. д.) и *литерато(-р, -ра* и т. д.). Отсюда ясно, что уже после первой попытки испытуемый получил полную информацию о следующей за *т* букве (если не *у*, то *о*), вторая же попытка угадывания была избыточной. В связи с этим в протоколе вместо первоначальных двух попыток была отмечена одна попытка. Аналогичные поправки вводились в протокол во всех случаях, когда число сделанных испытуемым попыток превышало их предельное количество, необходимое для отгадывания данной буквы.

В тех же случаях, когда экспериментатор или испытуемый сомневается в достоверности того или иного продолжения или степени его вероятности, вопрос может быть решен с помощью языковедов-экспертов. Помощь экспертов была использована при исследовании русского, румынского и английского материала.

Второй вид коррекции имеет целью проверить, насколько объективно выделил испытуемый достоверные продолжения. Эта коррекция должна также выявить те «нули информации», которые не были учтены в процессе эксперимента.

Выше уже указывалось (см. § 4), что к числу достоверных продолжений следует относить наряду с абсолютно достоверными и такие продолжения, которые обладают достаточно высокой вероятностью. Общие принципы выявления «нулей информации» рассмотрим на конкретном примере.

Предположим, что испытуемый, определив все буквы цепочки *△мать*, переходит к угадыванию следующей буквы. Предположим, что условная вероятность появления пробела после указанного букворяда (p) равна 0.999, а условная вероятность появления буквы *е* ($1-p$) — ср. имена собственные *Матье*, *Матьез* — соответственно равна 0.001.

В этом случае условная энтропия рассматриваемой буквенной позиции равна:

$$H = -p\log_2 p - (1-p)\log_2(1-p) = -0.999\log_2 0.999 - \\ - 0.001\log_2 0.001 = 0.0114 \text{ дв. ед.}$$

Если информант и экспериментатор объявят пробел после букворяда *△мать* достоверным продолжением, энтропия которого равна нулю, они сделают ошибку в 0.0114 дв. ед. Наоборот, признав, что рассматриваемая буквенная позиция дает два продолжения, они припишут ей в условиях эксперимента энтропию, равную одной дв. ед. В этом случае ошибка возрастет до 0.9886 дв. ед. и предсказание оказывается неэффективным. Отсюда ясно, почему к числу достоверных продолжений, отмечаемых цифровым символом нуль, должны быть отнесены не только абсолютно достоверные продолжения, но и такие продолжения, условная вероятность которых выше 0.889. В этом случае ошибка, возникающая при отнесении рассматриваемых буквенных позиций в число достоверных продолжений, будет меньше ошибки, появляющейся в том случае, если считать эти продолжения недостоверными.

Само собой разумеется, что в большинстве случаев степень достоверности продолжения может быть определена только исходя из лингвистической интуиции информанта, экспериментатора или экспертов.

Третий вид коррекции предусматривает расположение экспериментальных значений q_k^n в порядке убывания q и возрастания k, с тем чтобы приблизить опытное значение $\underline{H}_n$ к той оценке нижней границы энтропии, которая могла бы быть получена в результате идеального предсказания (см. § 10). Впрочем, как это видно из таблицы 3, указанный вид коррекции дает незначительное уменьшение величины $\underline{H}_n$.

§ 17. ОЦЕНКА ДОСТОВЕРНОСТИ РЕЗУЛЬТАТОВ

Достоверность результатов эксперимента по угадыванию может быть оценена путем их сравнения с соответствующими характеристиками, получаемыми на аналогичном языковом материале либо с помощью уже описанных методов, либо путем использования иных исследовательских приемов.

Первый вид сравнения использовался при оценке достоверности величины $\bar{H}_n''$ и $\underline{H}_n'$ в русских и английских беллетристических текстах. Расхождение наших оценок с аналогичными данными, полученными для русской прозы А. П. Савчук и для английской беллетристики К. Шенноном, а также Н. Бэртоном и Дж. Ликлайдером,

не превышает 10% [7, стр. 10—11; 25, стр. 129—130; 34; 46, стр. 652—653; 77, стр. 683—684].

Второй вид оценки предусматривал сопоставление результатов основного эксперимента с контрольными данными, полученными либо с помощью новых методов угадывания, либо путем расчета вероятностных спектров слов и буквенных цепочек.

При оценке достоверности основных результатов используется два новых лингво-психологических эксперимента: коллективное угадывание букв неизвестного текста и угадывание по методу Колмогорова.

Коллективное угадывание осуществляется по следующей программе. Достаточно большой группе испытуемых — носителей языка — предлагается предсказать букву, находящуюся на n-ом шаге связного текста или слова, причем $n-1$ букв этого отрывка или слова известны испытуемым. Каждый из испытуемых независимо от других записывает наиболее вероятное с его точки зрения буквенное продолжение для n-ого шага текста. После этого экспериментатор сообщает действительную букву и испытуемые переходят к следующему угадыванию.

Предполагается, что коллектив испытуемых предсказывает букву оптимальным образом. Поэтому на каждом шаге текста наибольшее число продолжений должно падать на наиболее вероятную букву, следующее количество предложений даст вторая по вероятности буква и т. д. Частное от деления числа испытуемых, предложивших данную букву, на общее число испытуемых можно рассматривать как условную вероятность p_k появления некоторой буквы k в данном участке текста или слова. Само собой разумеется, что коллектив испытуемых лишь условно может рассматриваться в качестве идеального предсказывающего устройства. В действительности реальные предсказания коллектива обычно довольно далеки от идеального угадывания. В частности, испытуемые обычно не называют редкие продолжения. Поэтому необходимо приблизить результаты угадывания к идеальному предсказыванию. Здесь можно предложить два приема. Во-первых, перед угадыванием буквы на n-ом шаге текста испытуемым сообщаются все возможные для этого шага буквенные продолжения (без указания их вероятностей). Во-вторых, не названные испытуемыми продолжения после завершения эксперимента вносятся в протокол и им приписываются сколь угодно малые вероятности (весь спектр вероятностей соответственным образом пересчитывается).

Протокол коллективного угадывания с только что указанной коррекцией показан в таблице 7.

Имея распределение условных вероятностей, можно получить, используя выражение

$$H_n' = -\sum_{k=1}^{S} p_k \log p_k, \qquad (41)\ [\text{ср. } (3)]$$

условную энтропию рассматриваемой буквенной позиции.

Возможна и более общая задача: можно, например, попытаться определить среднюю условную энтропию буквенных позиций в некоторой модели текста либо в слове определенной длины или морфемно-слогового строения. В этом случае коллективному угадыванию последовательно подвергается ряд отрывков или слов, составляющих достаточно репрезентативную выборку. Расчет средней условной энтропии осуществляется по формуле:

$$H_n'' = -\sum_{d_i^{n-1}} p\,(d_i^{n-1}) \sum_{k=1}^{S} p\,(j_{i,k} \mid d_i^{n-1}) \log p\,(j_{i,k} \mid d_i^{n-1}). \qquad (42)\ [\text{ср. } (4)]$$

Если угадываются отдельно взятые слова, то вероятность цепочки d_i^{n-1} может быть получена с помощью данных частотного словаря. В этом случае H_n'' близко к истинному значению H_n.

Если же эксперимент проводится на выборке связных текстов и вероятность $p\,(d_i^{n-1})$ определить невозможно, то приходится считать все цепочки d_i^{n-1} равновероятными. Расчет средней условной энтропии осуществляется здесь по формуле (41), причем, исходя из соображений, приведенных в § 8, нетрудно показать, что всегда

$$H_n' > H_n.$$

С помощью коллективного угадывания оценивалось распределение информации в русском, польском и румынском (молдавском) слове. Сопоставление результатов предсказания по полной и сокращенной программам с данными коллективного угадывания показало удовлетворительное схождение основных и контрольных результатов [24, стр. 113—115].

Таблица 7

Образец протокола и расчетная таблица протокола коллективного угадывания внеконтекстного слова *знаменитый*

Алфавит	З									
	Н	*А*	*М*	*Е*	*Н*	*И*	*Т*	*Ы*	*Й*	
а	13 0.220	52 0.961				6 0.109		2 0.036		
б										
в	2 0.035									
г	0.010									
д	10 0.170									
е	15 0.255		2 0.036	13 0.481			5 0.092	0.010	12 0.215	
ж										
з										
и	2 0.035					40 0.726				
й									42 0.755	
к			4 0.072							

Таблица 7 (продолжение)

Алфавит	З									
	Н	*А*	*М*	*Е*	*Н*	*И*	*Т*	*Ы*	*Й*	
л	0.010		0.010							
м	0.010		24 0.441						0.010	
н	10 0.170		16 0.295		54 1.000	0.010				
о	2 0.035	2 0.039		2 0.037		6 0.109		14 0.254		
п										
р	0.010									
с										
т			6 0.110				49 0.908			
у	0.010							0.010		
ф										
х									0.010	
ц										
ш										
щ						0.010				

Таблица 7 (продолжение)

Алфавит	H	A	M	E	H	Π	$Г$	$Ы$	$Н$
ы	0.010							36 0.654	
ь, ъ									
э									
ю	0.010		2 0.036						
я	0.010			13 0.482					0.010
Δ						2 0.036		2 0.036	
									54 1.000
H_n 4.220	2.891	0.238	2.075	1.491	0.000	1.338	0.433	1.382	0.983	0.000
	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}

Двоичные единицы информации

Примечание. В левом верхнем углу каждой клетки указано количество испытуемых, предложивших данное продолжение (всего в эксперименте участвовало 54 испытуемых). Ниже дана эмпирическая вероятность данного продолжения. Для тех продолжений, которые не были предложены информантами, даются их приблизительно взятая вероятность, равная 0,01. Энтропия первой буквы получена из расчета статистического спектра первых букв слов, взятых вне контекста.

Угадывание по методу Колмогорова осуществляется следующим образом. Испытуемому сообщается $n-1$ букв отрывка и предлагается последовательно угадывать n-ю, $(n+1)$-ю, . . ., $(n+\nu)$-ю буквы того же отрывка. При угадывании буквы, стоящей на соответствующем шаге текста, испытуемый должен дать один из следующих ответов:

а) назвать предлагаемую букву с большой степенью уверенности;

б) назвать предполагаемую букву с малой степенью уверенности;

в) назвать две возможные в данном участке отрывка буквы, каждая из которых обладает, по его мнению, одинаковой вероятностью;

г) назвать две возможные буквы, указав при этом, что одна из них более вероятна, чем другая;

д) отказаться от угадывания.

Называя букву с большой степенью уверенности, испытуемый записывает ее на пересечении столбца, соответствующего угадываемой букве, и первой строки. Если буква названа с малой степенью уверенности, она записывается во второй строке. Предложение о двух равновероятных буквах записывается в третьей строке, а предложение о неравновероятных буквах заносится в четвертую строку. Отказ от угадывания отмечается вопросительным знаком в первой строке. В нижней (пятой) строке записывается правильная буква, сообщаемая информанту после того, как он определит свое предсказание. Образец протокола показан в таблице 8.

Результаты угадывания обрабатываются по формуле:

$$\bar{H}(\nu+1) = -\frac{\gamma_n}{\nu+1}[f_0 \log f_0 + (1-f_0)\log(1-f_0)] -$$

$$-\frac{\gamma_n}{\nu+1}[f_1 \log f_1 + (1-f_1)\log(1-f_1)] -$$

$$-\frac{\gamma_n}{\nu+1}[f_2 \log f_2 + (1-f_2)\log(1-f_2)] + \frac{\gamma_6}{\nu+1} f_2 \log 2 -$$

$$-\frac{\gamma_r}{\nu+1}[f_3 \log f_3 + f_4 \log f_4 + (1-f_3-f_4)\log(1-f_3-f_4)] -$$

$$-\frac{\sum a_{r0} + \sum a_2 a_4 + \sum a_0 a_9}{\nu+1}, \qquad (43)$$

где $\bar{H}$ — средняя энтропия на букву на участке текста от n-ой до $(n+\nu)$-той буквы (общее число угадываемых букв

Таблица 8

Образец протокола угадывания слова *ребячьи* по методу Колмогорова

Строки	Шаги текста						
	n	$n+1$	$n+2$	$n+3$	$n+4$	$n+5$	$n+6$
1			?	[я]	[ч]		[и]
2		*а*					
3	*д/в*						
4						[ь] *л*	
5	*р*	*е*	*б*	*я*	*ч*	*ь*	*и*

Примечание. Правильные предсказания отмечены квадратом. Контрольному слову предшествует следующий контекст: *И вот, приехав в тот город, где друг мой в свое время справлял свадьбу, я встретил его жену и сына, который учится уже в пятом классе. И сын его меня расспрашивает, правда ли, что я служил с его отцом, на каких кораблях плавал, где бывал. Я отвечаю на вопросы так будто незнакомые ни первый взгляд. . .* («Известия», № 281, от 4 ноября 1965 г.).

равно $v+1$); γ_x — число предсказаний того или иного типа, например, γ_a — число предсказаний с большой степенью уверенности (правильных и неправильных), $\gamma_б$ — число предсказаний с малой степенью уверенности (правильных и неправильных) и т. д.

Символы f имеют следующие значения:

f_0 — вероятность не угадать в случае предсказания с большой степенью уверенности;

f_1 — вероятность не угадать в случае предсказания одной буквы с малой степенью уверенности;

$1-f_2$ — вероятность не угадать в случае, когда испытуемый предлагает две равновероятных буквы;

f_3 — вероятность угадать наиболее вероятную букву из двух предлагаемых испытуемым неравновероятных букв;

f_4 — вероятность угадать из двух неравновероятных букв маловероятную букву.

Значения f получены следующим образом. Для каждого из перечисленных выше случаев угадывания возможны такие исходы:

а) a_1 — правильное предсказание, a_2 — неправильное предсказание;

б) a_3 — правильное предсказание, a_4 — неправильное предсказание;

в) a_5 — правильное предсказание, a_6 — неправильное предсказание;

г) a_7 — правильное предсказание для более вероятной буквы, a_8 — правильное предсказание для менее вероятной буквы, a_9 — неправильное предсказание;

д) a_{10} — отказ от предсказания.

Условимся, что число исходов типа a_i равно a_i, тогда

$$f_0=\frac{a_2}{a_1+a_2};\quad f_1=\frac{a_4}{a_3+a_4};\quad f_2=\frac{a_5}{a_5+a_6};\quad f_3=\frac{a_7}{a_7+a_8+a_9}.$$

Далее будем считать, что

$$f_4=\frac{a_8}{a_7+a_8+a_9}.$$

Σa_{10} — сумма отказов, $\Sigma a_2 a_3$ — сумма ошибок в случаях «а» и «б», $\Sigma a_6 a_9$ — сумма ошибок в случаях «в» и «г».[12]

Схема Колмогорова позволяет получать значения энтропии, приближающиеся к истинному среднему значению в блоке из $v+1$ букв для данного участка текста. С помощью схемы Колмогорова Н. В. Петровой была получена оценка энтропии французского языка, которая попадала в интервал между верхней и нижней оценками, полученными по методам, описанным в §§ 9—12. Аналогичный результат для русского языка, полученный в группе А. Н. Колмогорова также оказался заключенным между верхней и нижней оценками, полученными при обработке результатов полной программы угадывания.

* * *

Суммарные значения энтропии на слово, полученные из основного эксперимента по угадыванию, могут быть сопоставлены с аналогичными оценками, добытыми путем расчета распределений вероятностей слов в тексте (частотные словари). Этот прием был использован для оценки достоверности информационных характеристик русского,

[12] Описание схемы Колмогорова заимствовано из работ [23, 60].

английского и румынского слова [ср. 1; 3, стр. 178; 4; 13; 24, стр. 120—121; 68а; 77, стр. 671—673].

При определении достоверности H_∞ и $\underline{H}_\infty$ может быть использована также полученная путем расчета дальних корреляционных связей оценка H_∞ [35, стр. 111—117; ср. также 12, стр. 462—463].

При сопоставлении числовых результатов основного и контрольного эксперимента, а также данных, полученных путем применения полного и сокращенного эксперимента, используются обычно среднее арифметическое и среднее квадратическое отклонения. При этом данные полного эксперимента рассматриваются как теоретические величины. В работе [8, стр. 7—8] сделана попытка применить для этих целей нормированный критерий χ^2, а также критерий знаков.

Глава II

ИНФОРМАЦИОННЫЕ ХАРАКТЕРИСТИКИ ТЕКСТА

§ 16. ИНФОРМАЦИЯ И ИЗБЫТОЧНОСТЬ В ЕВРОПЕЙСКИХ И НЕЕВРОПЕЙСКИХ ЯЗЫКАХ

Выше было показано (см. § 1), что, снимая энтропию того или иного участка текста, мы сообщаем равное ему по абсолютной величине количество информации [ср. 16, стр. 3]. В связи с этим количественную оценку лингвистическим единицам можно давать не в терминах энтропии (H), а через аналогичные им величины информации (I) — ср. стр. 8, примеч. 4.

По описанной в §§ 2—13 программе были получены значения предельной информации ($I_\infty = H_\infty$) и избыточности (R) для пяти европейских [см. 7, 25, 68а, 69] и двух тюркских языков (см. 2) (для тюркских языков эти данные имеют пока предварительный характер).

Значения I_∞, $\underline{I}_\infty$ по русскому языку получены как средние арифметические соответствующих значений для 30-й, 31-й, 32-й, . . . 100-й букв, для английского, польского и казахского языков — как среднее арифметическое I_n, $\underline{I}_n$ для 30-й, 40-й, 50-й, 75-й, 100-й, 150-й и 200-й букв, для французского — как среднее арифметическое их значений 30-й, 40-й, 50-й, 60-й, 75-й, 100-й букв, для румынского — как среднее арифметическое для 30-й, 40-й, 50-й, 75-й, 100-й букв и, наконец, для азербайджанского — как среднее арифметическое для 30-й, 40-й, 50-й, 60-й, 70-й, 80-й, 90-й, 100-й, 150-й, 200-й букв. По этой же методике в настоящее время исследуется энтропия новогреческого и узбекского языков. Кроме того, путем коллективного угадывания получена оценка энтропии испанского языка в его кубинском варианте.

В свое время было показано, что, начиная с двадцатой-тридцатой букв текста, значения I практически перестают

зависеть от n [ср. 25, стр. 121 и 124; 46]. Возможно, это связано с объемом непосредственной (оперативной) памяти человека [65]. Поэтому тридцатая буквенная позиция рассматривается в качестве левой границы отрезка, по которому определяются оценки I_∞. Использование буквенных позиций, лежащих ближе к началу текста, дает явное завышение значений $\bar{I}_\infty$ и $\underline{I}_\infty$. Об этом можно судить, в частности, по данным таблицы 3.

Из значений $\bar{I}_\infty$ и $\underline{I}_\infty$ могут быть соответственно получены нижняя ($\underline{R}$) и верхняя ($\bar{R}$) оценки избыточности языка. При этом используется выражение

$$R = \frac{I_0 - I_\infty}{I_0} . \qquad (44)$$

Значения $\bar{I}_\infty$, $\underline{I}_\infty$, $\underline{R}$ и $\bar{R}$ для отдельных стилей и языка в целом показаны в таблице 9.

Нетрудно заметить, что все семь языков обнаруживают большую близость в общих оценках энтропии и избыточности. Как уже говорилось (см. § 15), эти оценки для языка в целом получены путем анализа достаточно большого объема материала (исключение представляет лишь казахский материал).

Коэффициент вариации $v = \frac{\sigma}{I} 100\%$ для значений I_∞ и $\underline{I}_\infty$ относительно языка в целом сравнительно невелик (в среднем он равен примерно 23%). Контрольные значения $\bar{I}_\infty$, $\underline{I}_\infty$ и I_∞, заимствованные из работ других авторов и полученные в результате проверочного эксперимента (см. таблицу 10), близки к нашим данным. Все это дает право считать верхние и нижние оценки энтропии, и избыточности, приводимые в таблице 9, достаточно надежными. Иными словами, избыточность развитого языка находится, очевидно, в пределах между 70 и 85%.

Что касается отдельных стилей, то материала по ним было мало (33 текста на один стиль). Поэтому статистический разброс здесь значительно выше, чем для языка в целом (в среднем $v = 32.3\%$, а для пиковых его значений достигает 60%). Расхождения между разговорными и беллетристическими текстами настолько малы, что несомненно перекрываются случайным разбросом экспериментальных данных. При этом следует помнить, что статистическому анализу были подвергнуты не сами разговорные тексты, а лишь их письменная запись. В этом случае не могла быть

Таблица 9

Сравнительные данные об энтропии (в дв. ед.) и избыточности (в %) пяти европейских и двух тюркских языков

Функциональный стиль («подъязык»)	Русский язык. Таблица с учетом поправки Пуа вычисления, см. § 16 [25]			Польский язык [11]			Английский язык [8]			Французский язык [25]		
	I_∞	$\underline{I}_\infty$	$\underline{R}$	$\bar{I}_\infty$	$\underline{I}_\infty$	$\underline{R}$	$\bar{I}_\infty$	$\underline{I}_\infty$	$\underline{R}$	$\bar{I}_\infty$	$\underline{I}_\infty$	$\underline{R}$
Разговорная речь	1.40	0.83	72.0	1.18	0.69	76.3	1.47	0.90	69.4	1.32	0.81	68.5
Беллетристика	1.19	0.70	76.3	1.29	0.83	74.5	1.10	0.65	77.1	1.36	0.78	71.0
Деловая речь	0.83	0.49	83.4	0.83	0.53	83.6	0.82	0.37	82.9	0.77	0.45	83.9
Язык в целом	1.37	0.82	72.1	1.28	0.76	74.7	1.35	0.74	71.9	1.38	0.79	70.6

Функциональный стиль («подъязык»)	Румынский язык			Азербайджанский язык [2]			Казахский язык [2]			$\bar{R}$
	I_∞	$\underline{I}_\infty$	$\underline{R}$	$\bar{I}_\infty$	$\underline{I}_\infty$	$\underline{R}$	$\bar{I}_\infty$	$\underline{I}_\infty$	$\underline{R}$	
Разговорная речь	1.24	0.68	74.2	85.9	—	—	—	—	—	82.9
Беллетристика	1.26	0.78	73.8	83.8	—	—	—	—	—	83.6
Деловая речь	1.23	0.68	74.4	85.9	—	—	—	—	—	90.4
Язык в целом	1.34	0.72	72.4	85.0	1.77	1.07	65.2	79.0	1.59	1.07	70.9	81.0

Продолжение

Таблица 10

Оценка достоверности данных об информации пяти европейских языков [1]

Функциональный стиль («подъязык»)	Русский язык			Польский язык [11]			Английский язык			Французский язык			Румынский (молдавский) язык		
	Угадывание по полной программе		Эксперимент Колмогорова ($\approx I_\infty$)	Угадывание по полной программе		Коллективное угадывание $H'_n \approx I_n$	Угадывание по полной программе		Расчет дальних корреляций между символами текста ($\approx I_\infty$)	Угадывание по полной программе		Эксперимент Колмогорова $\approx I_\infty$	Угадывание по полной программе		Коллективное угадывание
	I_∞	I_∞		I_∞	I_∞		I_∞	I_∞		I_∞	I_∞		I_∞	I_∞	
Беллетристика	1.49 (1.19)	0.74 [II] (0.70)	1.10 [III]	--	—	—	а) 1.45 б) 1.36 (1.10)	0.74 [IV] 0.79 [V] (0.65)	0.94 [VI]	--	--	—	—	—	--
Язык в целом	—	--	—	(1.28)	(0.76)	(0.95)	—	--	—	(1.38)	(0.79)	1.00 [VII]	(1.34)	(0.72)	0.725 [VIII]

[1] В скобках указаны данные, полученные в группе «Статисти ка речи» по полной и сокращенной программам (см. таблицу 9). Цифры без скобок показывают контрольные значения I_∞, I_∞ и I_∞, полученные другими авторами (ср. сноски). I_∞ и I_∞ получены пу тем применения поправок к величинам I''_∞ (достоверные продолже ния) и I'_∞ (по упорядоченным q_k), см. §§ 9, 16. Везде даны двоич ные единицы информации. [II] По данным А. П. Савчук [см. 25, стр. 130]. [III] Информация А. Н. Колмогорова. [IV] [77, стр. 686]. [V] [46, стр. 652—653]. [VI] [18, стр. 29]. [VII] [35, стр. 115]. [VIII] По данным И. В. Петровой.

Таблица 11

Информация (в дв. ед) и избыточность (в %%) стилей адыгейского, армянского и чешского языков сравнительно с аналогичными данными по русскому языку [1]

	Адыгейский язык [19, стр. 165—196]		Армянский язык [15, стр. 168—172]		Чешский язык [51, стр. 155] [II]			
	$\delta\frac{I'_{16}+I'_{21}+I'_{41}}{3}$	R	$\delta\frac{I'_{15}+I'_{16}+\ldots+I'_{20}}{6}$	$\delta\frac{I_{15}+I_{16}+\ldots+I_{20}}{6}$	R	$\bar{R}$	$\delta I'_{30}$	I'_{30}	R	$\bar{R}$
Беллетристика	2.26 (1.52)	56.0 (68.2)	1.38 (1.58)	0.78 (1.13)	73.9 (68.4)	85.3 (77.4)	1.36 (0.81)	1.08 (0.50)	74.4 (83.8)	80.0 (90.0)
Деловые тексты	--	--	1.36 (1.15)	0.71 (0.90)	73.8 (77.2)	86.5 (82.0)	1.41 (1.31)	1.19 (0.81)	73.8 (73.8)	77.9 (83.8)

[1] Цифры без скобок указывают значения энтропии и избыточ ности для буквенной позиции или участка текста по одному из трех заглавных языков. Цифры в скобках показывают соответству ющие данные по русскому языку [см. 25, стр. 128]. Поскольку верхние оценки энтропии в адыгейских, армянских и чешских тек стах рассчитывались без учета достоверных продолжений, значения I' умножаются на эмпирический поправочный коэффициент δ, с по мощью которого эти последние приводятся к значениям I. [II] Дан ные по беллетристическим текстам получены путем обобщения ре зультатов по выборкам A, C, E [51, стр. 166], по выборкам F, G, а данные по деловому стилю взяты из выборок A, C, E [51, стр. 166].

учтена та информация, которая передается обстановкой беседы, интонацией, мимикой и другими коммуникативными средствами, характеризующими устное общение. Если бы удалось учесть всю эту информацию, то оценки избыточности устной речи были бы значительно выше (а энтропийные оценки соответственно ниже).

Что касается несколько большей по сравнению с беллетристикой избыточности русских, польских, английских и французских деловых текстов, то она, очевидно, отражает реальное положение вещей.[13]

Следует также всегда помнить, что значения I_∞ и $\underline{I}_\infty$ очень чувствительны по отношению к методике проведения эксперимента. Значительные расхождения в оценках дают адыгейский, армянский и чешский языки (см. таблицу 11). Напротив, пять европейских языков показывают большую близость. Это объясняется, очевидно, тем, что в последнем случае угадывание проводилось по одной строго регламентированной методике. Что же касается эксперимента по адыгейскому, армянскому и чешскому языкам, то здесь, очевидно, применялись несколько иные критерии при выборе информанта и использовании справочно-вспомогательного аппарата. В частности, отсутствием этого аппарата объясняется, очевидно, исключительно высокое значение I_∞ в адыгейских текстах [19, стр. 165—166].

§ 19. ОБЩАЯ ИНФОРМАЦИОННАЯ СХЕМА ТЕКСТА

Письменные тексты любого языка состоят из разного вида четко отграниченных друг от друга дискретных единиц. В языках, использующих буквенный алфавит, такими единицами являются: буква, цепочка букв, ограниченная слева и справа пробелами (словоформа), цепочка словоформ, ограниченная слева и справа точками (предложение), цепочка предложений, ограниченная слева пустыми участками строк (абзац). В качестве единиц следующих иерархических уровней членения текста выступают параграфы, главы, разделы, части и т. д.

[13] В ряде случаев избыточность деловой речи достигает 95—96%. Этим путем гарантируется, например, надежность в приеме команд аэродрома экипажем идущего на посадку самолета [см. 54, стр. 245—251; 55, стр. 595—596].

Единицы более высокого уровня, чем буквы, обычно имеют внутреннее лингвистическое членение. В словоформах выделяются слоги и морфемы, в предложениях — словосочетания и синтагмы. Однако это членение не имеет обычно однозначного и недвусмысленного выражения.

Сочетаемость единиц каждого уровня регулируется двумя типами ограничений: ограничениями, заложенными в самой структуре языкового кода, и экстралингвистическими ограничениями, присущими выражаемому данным текстом содержанием.

Ограничения первого типа, почти целиком или в значительной мере регулирующие сочетаемость букв, слогов, морфем, постепенно заменяются при переходе на более высокие уровни ограничениями второго — экстралингвистического типа.

Исследование линейной структуры текста, в том числе с помощью метода угадывания, показывает, что по мере продвижения от начала текста соотношение обоих типов ограничений меняется. В свое время было обнаружено, что уже на пятом-шестом шаге текста комбинаторика букв и слогов подавляется ограничениями, связанными с сочетаемостью морфем. Начиная примерно с восьмой-десятой буквенной позиции, комбинаторика морфем ограничивается сочетаемостью слов. Затем, по мере развертывания текста на комбинаторику слов напластовываются ограничения в сочетаемости словосочетаний и предложений, а на эти последние в свою очередь накладываются ограничения, связанные, в частности, с комбинациями параграфов, глав, частей книги или статьи. Нетрудно заметить, что в начале текста доминируют ограничения, заложенные в лингвистическом коде. По мере развертывания текста на них напластовываются экстралингвистические ограничения [24, стр. 117].

Одна из основных задач информационно-статистических исследований состоит в том, чтобы оценить эти ограничения не только в их совокупности, но и по отдельности. Тот факт, что наш эксперимент ставится лишь относительно начального участка текста (до сотой или двухсотой буквы), позволяет исследовать лишь те ограничения, которые касаются сочетаемости букв, слогов, морфем, слов, словосочетаний и лишь отчасти предложений.

Начнем с того, что опишем методику оценки этих ограничений в их совокупности.

Рассматривая цепочки частных значений I (соответственно I'_n или I_n), нетрудно убедиться, что они обнаруживают тенденцию к убыванию в зависимости от роста значений n. Представим каждую из этих цепочек не как последовательности дискретных значений энтропии, но как непрерывно кривые (точнее — как непрерывные функции непрерывного аргумента). Если отвлечься на время от непериодических колебаний, относя их за счет разброса, то эти цепочки значений I могут быть апроксимированы в первом приближении теоретической показательной кривой (экспонентной) вида

$$I_\xi^T = (I_0 - I_\infty)\, e^{-s\xi} + I_\infty, \qquad (45)$$

где $I_\xi^{(T)}$ — верхний или нижний теоретические пределы информации в данном участке текста; ξ — непрерывный аргумент функции, заменяющий дискретные величины n; I_0 — информация алфавита, а I_∞ — информация языка или стиля (мы будем иметь каждый раз дело с верхней или нижней оценкой этого предела), см. §§ 9, 12; e — основание натуральных логарифмов; s — специально рассчитываемый для каждой кривой контекстный коэффициент.[14]

Убывание значений I_n (I_ξ) по мере удаления от начала текста легко объяснимо с лингвистической точки зрения. Действительно, если бы в лингвистическом коде (языке) не были бы заложены ограничения на сочетаемость букв (фонем), слогов, морфем и т. д. (ср. выше), то неопределенность в каждом участке текста была бы равна энтропии алфавита (I_0). Тот факт, что конкретные значения I_n гораздо меньше величин I_0, объясняется тем, что на каждую буквенную позицию n накладывается некоторая совокупность из перечисленных выше лингвистических ограничений. Эта совокупность ограничений, которую мы будем называть к о н т е к с т н о й о б у с л о в л е н н о с т ь ю — $K_\xi(K_n)$, может быть оценена количественно как разность $I_0 - I_n$. Заменив I_n правой частью формулы (45) и произведя некоторые упрощающие преобразования, получим общее выражение контекстной обусловленности в данном участке текста

$$K_\xi = (I_0 - I_\infty)(1 - e^{-s\xi}). \qquad (46)$$

По существу для каждого языка или стиля мы имеем по две формулы относительно выражений (45) и (46). Одну для I_ξ (соответственно K_ξ), другую — для $I_\xi(\underline{K}_\xi)$ или $I'_\xi(\underline{K}'_\xi)$.

Выражение (46) служит показателем того, как нарастают информационные связи между единицами текста по мере удаления от его начала. График этой зависимости относительно верхних оценок энтропии для русского языка в целом и его делового стиля см. на рис. 2. Предельная контекстная обусловленность ($K_\infty = I_0 - I_\infty$) для пяти языков показана в таблице 12.

Таблица 12

Предельная контекстная обусловленность по оценкам I для пяти европейских языков (в дв. ед.)

	Русский язык	Польский язык	Английский язык	Французский язык	Румынский язык
Разговорная речь	3.60	3.96	3.29	3.44	3.52
Беллетристика	3.81	3.75	3.66	3.40	3.50
Деловые тексты	4.17	4.21	3.94	3.99	3.53
Язык в целом	3.63	3.76	3.41	3.38	3.42

Показателем хода зависимости K_ξ (равно как и I_ξ) является к о н т е к с т н ы й к о э ф ф и ц и е н т s. Чем больше величина s, тем скорее идет увеличение значений K_ξ (соответственно уменьшение I_ξ). Иначе говоря, коэффициент s является показателем темпа роста коллективных связей. Характерно, что наибольшую величину s дают тексты делового стиля (см. таблицу 13, рис. 2), в которых благодаря использованию большого количества устойчивых словосочетаний и ограниченного круга лексики контекстные связи языковых единиц устанавливаются быстрее, чем в текстах других стилей.

[14] Показательная кривая $I^{(T)}$ рассчитывается на участке с 31-й по 100-ю букву с помощью метода наименьших квадратов. На участке от 1-й до 30-й буквы этот метод не дает минимализации линейных (и квадратичных) отклонений. Поэтому начальную часть графика целесообразно получать путем подбора с таким расчетом, чтобы сумма линейных отклонений экспериментальных точек от экспоненты стремилась к нулю [ср. 25, стр. 121, примеч. 17].

Таблица 13

Значения контекстного коэффициента s

	Русский язык		Французский язык	
	верхняя граница	нижняя граница	верхняя граница	нижняя граница
Разговорная речь	0.22	0.31	0.32	0.31
Беллетристика	0.21	0.29	0.26	0.29
Деловые тексты	0.24	0.32	0.34	0.42
Язык в целом	0.19	0.31	0.30	0.36

Лингвистическая природа коэффициента s достаточно сложна. Самый начальный ход экспоненты от $n=0$ до $n=1$ обусловлен введением статистических ограничений на

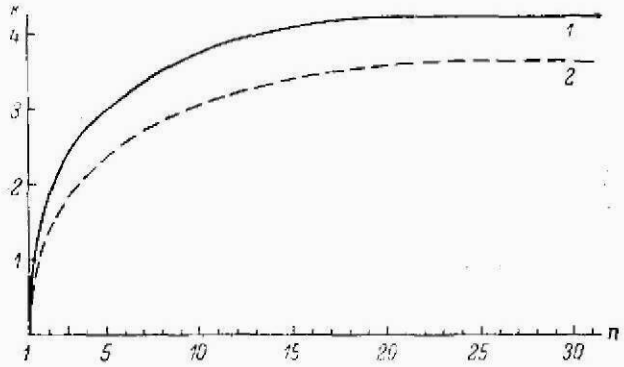

Рис. 2. Рост контекстной обусловленности в деловых текстах и в смешанной выборке текстов русского языка.

Условные обозначения: *1* — деловой текст; *2* — смешанный текст.

употребление первых букв слова. Затем, при n порядка нескольких единиц ход кривой определен закономерностями в сочетаемости букв (фонем), слогов и отчасти морфем. Позднее вступают в действие статистические ограничения в сочетаемости слов с их грамматическими формами, затем, как уже говорилось, появляются ограничения, связанные с комбинаторикой более крупных единиц. Таким образом, коэффициент s следует рассматривать как суммарный показатель различных парадигматических и синтагматических вероятностных ограничений. Этот коэффициент по всей вероятности достаточно хорошо отражает действие буквенных (фонемных), слоговых, морфемных сочетаемостей, а также сочетаемости слов. Что касается синтактико-композиционных связей, действие которых распространяется на отрезки текста длиной в несколько десятков букв (фонем), то эти связи плохо описываются контекстным коэффициентом. Ведь для порядка 30 и более букв показатель s дает ничтожное изменение в ходе экспоненты.

* * *

В группе «Статистика речи» (ср. выше, стр. 42, примеч. 11) разрабатываются и другие приемы аппроксимации распределения информации в тексте.

Г. П. Богуславская [8] предложила для этого формулу:

$$I_n^T = H_1\left[1 - a\left(1 - \frac{1}{n}\right)\right], \qquad (47)$$

где I_n^T — теоретическое значение информации (по верхней или нижней границе) для n-ой буквенной позиции текста; H_1 — энтропия первой буквы текста; a — получаемый эмпирическим путем постоянный коэффициент, аналогичный нашему коэффициенту s.

* * *

В. М. Калинин предложил процедуру расчета информации для каждой буквенной позиции текста, исходя из распределения вероятностей употребления слов различной длины и информации (энтропии), падающей на данную букву слова определенной длины. Здесь используются следующие рассуждения.

Пусть $H_\lambda(m)$ — верхняя или нижняя оценка энтропии m-той буквы внетекстового слова, которое состоит из λ букв, включая пробел; P_λ — вероятность того, что наугад взятое слово из текста имеет вместе с пробелом λ букв. Данные о величинах P_λ и $H_\lambda(m)$ относительно русских внетекстовых слов приведены в таблице 14. Предположим, что в пределах предложения вероятность того, что слово имеет определенную длину, пренебрежимо мало зависит от номера буквы в предложении. Тогда, считая, что мак-

симальная длина слова в данном языке равна t, имеем выражение

$$I(m) = \sum_{\lambda=m}^{t} P_\lambda H_\lambda(m),$$

показывающее, сколько информации содержится в m-той букве слов текста.

Информация, содержащаяся в первой буквенной позиции предложения, будет по определению математического ожидания равна

$$I_1 = \sum_{\lambda=2}^{t} P_\lambda H_\lambda(1) = I(1).$$

Аналогичным образом

$$I_2 = \sum_{\lambda=2}^{t} P_\lambda H_\lambda(2) = I(2).$$

Третья буква предложения может быть либо третьей буквой слова, состоящего из 3-х, 4-х, . . . t букв, либо первой буквой слова, состоящего из 2-х, 3-х, . . . t букв, при условии, что первое слово предложения состояло всего из двух букв (т. е. одной буквы + пробел). Информация на третьей буквенной позиции предложения равна

$$I_3 = \sum_{\lambda=3}^{t} P_\lambda H_\lambda(3) + P_2 \sum_{\lambda=2}^{t} P_\lambda H_\lambda(1) = I(3) + P_2 I(1). \qquad (48)$$

Аналогично

$$I_4 = I(4) + P_2 I(2) + P_3 I(1), \qquad (49)$$

$$I_5 = I(5) + P_2 I(3) + P_3 I(2) + P_4 I(1) + P_2^2 I(1). \qquad (50)$$

По этой методике Л. А. Гусакова подсчитала распределение информации по верхней границе в первых двенадцати буквах русского текста. Как видно из таблицы 14, это распределение достаточно близко к распределению значений $\hat{H}_n = \hat{I}_n$, полученных из эксперимента по угадыванию разговорных текстов.

Методика, предложенная В. М. Калининым, может быть улучшена путем введения в выражения (48—50) и особенно в соответствующие им формулы расчета $\hat{I}_n$ при $n \geqslant 5$ некоторого поправочного коэффициента, отражающего нарастание лексико-грамматических связей в тексте (ср. ниже § 21). Кстати, близость значений I_n и $\hat{I}_n$, о которой мы только что говорили, вероятно, объясняется тем, что степень лексико-грамматической связанности слов в разговорном тексте ниже, чем в текстах других жанров, где расхождение значений I_n и $\hat{I}_n$ будет гораздо более заметным (ср. данные таблицы 14 и таблиц 2, 3, 5, помещенных в работе [25]).

Таблица 14

Распределения информации (дв. ед.) в усредненном двенадцатибуквенном тексте (русский язык), полученные путем использования методики Калинина ($\hat{I}$) и с помощью угадывания (I)

	1	2	3	4	5	6	7	8	9	10	11	12
$\hat{I}$	4.22	2.73	2.75	2.16	1.76	1.61	1.92	1.86	1.97	2.04	2.04	1.97
I (разговорная)	4.22	2.63	2.50	2.14	1.61	1.76	1.75	2.31	2.21	2.63	2.25	1.99

§ 20. ЛЕКСИЧЕСКАЯ СХЕМА ТЕКСТА

Получив суммарные оценки и общий ход развития статистической информации текста, рассмотрим теперь ее распределение внутри текста, а также внутри его отдельных единиц. Для этого, используя результаты индивидуального угадывания, попытаемся построить такую усредненную информационную схему, которая отражала бы членение текста на слова (эту схему мы будем называть лексической схемой текста).

Чтобы построить такую схему, необходимо, во-первых, определить средние длины первого, второго и т. д. слов в текстах данного языка или стиля, во-вторых, получить схемы распределения информации для первого, второго, третьего и т. д. слов усредненного текста. Первая задача решается обычным статистическим путем, вторая требует особой группировки того материала, который был получен из эксперимента по угадыванию букв. Принципы этой группировки иллюстрирует таблица 15, представляющая собой участок матрицы, по которой рассчитывалась лексическая схема русского текста. В первом столбце со-

Таблица 15

Участок матрицы, использующейся для расчета лексической схемы текста

№№ текстов	Столбцы											
	1	2	3	4	5	6	7	8	9	10	11	12
	1-я буква слова I	2-я буква слова I	середина слова I	предпоследняя буква слова I	последняя буква слова I	пробел	1-я буква слова II	3-я буква слова II	середина слова	предпоследняя буква слова II	последняя буква слова II	пробел
1	*м*	*и*	*ог*	*и*	*е*	△	*д*	*е*	*лсея*	*т*	*ы*	△
	8	*3*		*1*	*1*	*0*	*5*	*3*		*1*	*0*	*0*
2	*д*	—	—	—	*о*	△	*е*	*с*	*чс*	*р*	*а*	△
	9	—			*3*	*2*	*6*	*2*		*1*	*1*	*0*
3	*я*	—	—	—	—	△	*п*	*е*	-	—	*л*	△
	11	—		—	—	*1*	*1*	*2*			*1*	*1*

браны первые буквы слов из каждого текста. Во втором столбце указываются вторые буквы первых слов при условии, что эти слова имеют длину не менее трех букв. Для одно- и двухбуквенных слов во втором столбце ставится прочерк. В четвертый столбец попадают предпоследние буквы слов длиной в четыре, пять и т. д. букв. Для коротких слов в четвертом столбце ставится прочерк. В пятый столбец помещаются последние буквы слов, имеющих длину две и более букв.[15] В шестой столбец записываются все пробелы между первым и вторым словами. При каждой букве и пробеле указывается количество попыток, понадобившихся для их угадывания. В третий столбец заносятся середины длинных слов без указания количества попыток. Таким же образом расписывается второе и последующие слова текста.

Аналогичное построение имеет матрица для расчета лексической схемы текста, отгадывавшегося по сокращенной программе. Ее отличие от первой матрицы состоит

[15] Нетрудно заметить, что число букв, попадающих в столбец, который обобщает последние буквы слов (соответственно вторые или предпоследние) может быть меньше общего количества текстов, представленных в матрице.

лишь в том, что угадывание буквы с двух и более попыток отмечается одним цифровым символом *2* (см. § 5).

Только что описанная матрица содержит распределения числа попыток, понадобившихся для угадывания первых, вторых, предпоследних и последних букв первого, второго, третьего и т. д. слов текста, а также аналогичные распределения относительно пробелов между словами.

Обработка этих распределений по формулам (17, 22 или 26) дает верхние и нижние оценки информации (энтропии) для первых двух и для последних двух букв слов нашего усредненного текста, а также оценки информации на пробелах между этими словами. Информация, падающая на середины слов (ср. столбец 3 в таблице 15) оценивается как среднее арифметическое от информаций, приходящихся на первую, вторую, пред-

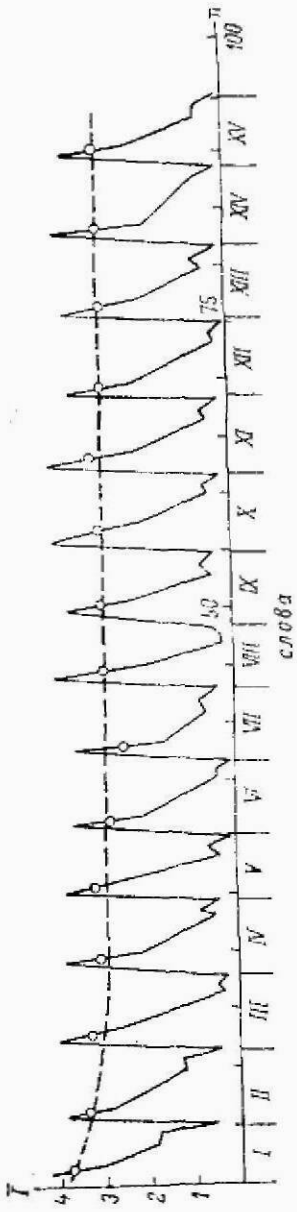

Рис. 3. Лексическая схема усредненного русского текста.

Условные обозначения: *1* — полигон распределения информации по верхней границе с учетом достоверных продолжений; *2* — значения $\frac{F_1^1+F_2^2}{2}$; *3* — графих значений $\frac{F_1^2+F_2^3}{2}$.

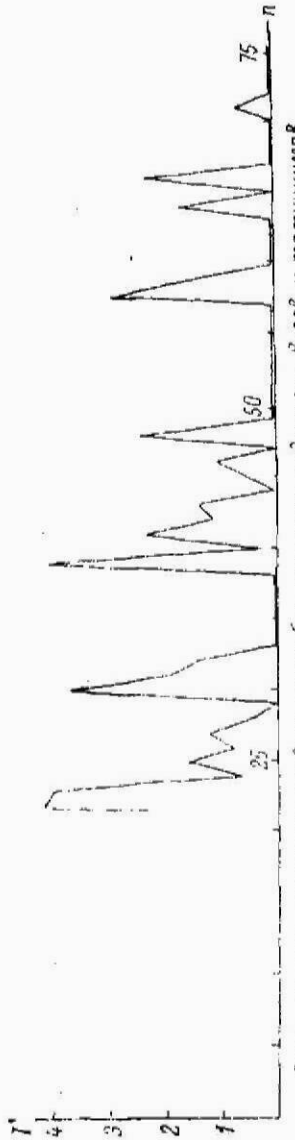

Рис. 4. Лексическая схема конкретного русского текста. При построении графика использовались данные коллективного угадывания газетного текста (газ. «Комсомольская Правда» от 20-го августа 1966 г.), при условии, что испытуемым были сообщены первые три слова (21 буква) отрывка.

последнюю и последнюю буквы усредненного слова. В итоге мы получаем схему распределения информации в тексте, учитывающую его членение на слова. Полученная путем обработки результатов угадывания по полной программе схема русского текста длиной в 15 слов показана в таблице 16 и на рис. 3. Аналогичные схемы были получены Г. П. Богуславской [8, стр. 17] по сокращенной программе угадывания для английского и румынского языков [8, стр. 17; 68а]. По французскому языку Н. В. Петровой были получены значения I' лишь для первой и второй букв в словах текста [69, стр. 142, 150—151].

Обработка результатов коллективного угадывания текста по формуле (41) (ср. § 17, таблицу 7) дает лексическую информационную схему конкретного текста (см. рис. 4). Используя описанные выше приемы, можно суммировать статистически достаточное число таких частных схем и получить новую усредненную лексическую схему, которая может быть использована для оценки достоверности той схемы текста, которая была построена путем обработки результатов индивидуального угадывания.

В заключение отметим, что, как показывают данные русского, английского и отчасти французского, испанского и румынского языков (см. ниже, § 23), распределение информации в тексте имеет квантовый характер. Начала слов несут максимумы информации, в то время как последние буквы слов и особенно следующие за ними пробелы оказываются либо мало информативными, либо вообще избыточными. Квантовый характер распределения статистической информации связан, очевидно, с теми особенностями, которые характеризуют работу головного мозга человека в ходе переработки им лингвистического текста.

§ 21. ЛЕКСИЧЕСКАЯ ОБУСЛОВЛЕННОСТЬ ЕДИНИЦ ТЕКСТА

В ходе эксперимента было замечено, что по мере продвижения от начала текста испытываемый все чаще угадывает вторую, а иногда и первую букву слова, опираясь не на комбинаторику букв, а на предшествующий лексический контекст.[16] Это дало возможность предположить, что, исследуя убывание сумм информаций, падающих на первую и вторую буквы слов (ср. рис. 3), можно оценить нарастание лексических связей в тексте. Нарастание этих связей, которое мы будем обозначать как рост лексической обусловленности (L), было оценено в первом приближении относительно французского языка [см. 69, стр. 142] с помощью показательной кривой, имеющей вид:

$$L_i = (I_1^{(a)} - I_\infty^{(a)})\,(1 - e^{-l i}). \qquad (51)$$

где $I_1^{(a)}$ — среднее арифметическое информаций, вычисленных для верхней или нижней границ, которые падают на 1-ю и 2-ю буквы первого слова текста, $I_\infty^{(a)}$ — предел лексической обусловленности текста (ср. величины I_∞ в выражениях (45) и (46)), l — лексический коэффициент,

[16] Первая буква слова, даже в тех случаях, когда испытуемому известен достаточно длинный предшествующий контекст, чаще всего угадывается по таблице частотности начальных букв слова.

характеризующий темп нарастания лексических связей в тексте (ср. коэффициент s в формулах (45) и (46)), остальные обозначения имеют тот же смысл, что и в предшествующих выражениях.

Таблица 16

Распределение информации (дв. ед.) в лексической схеме текста *по верхней границе информации*

№№ слов и их средние длины (цифры в скобках) без учета пробела	Количество информации, приходящееся на		$\frac{I_1^0 + I_2^0}{2}$	Количество информации, приходящееся на		
	первую букву слова (I_1^0)	вторую букву слова (I_2^0)		предпоследнюю букву слова	последнюю букву слова	пробел
I (4.40)	4.22	3.12	3.67	1.87	1.83	0.55
II (5.50)	3.77	3.00	3.38	1.07	1.23	0.40
III (5.67)	3.98	2.66	3.32	0.41	0.47	0.25
IV (5.16)	3.92	2.14	3.02	0.51	0.81	0.45
V (5.60)	3.79	2.49	3.14	0.29	0.61	0.13
VI (5.90)	3.58	1.98	2.78	0.35	0.39	0.06
VII (5.52)	3.43	1.43	2.43	0.57	0.66	0.35
VIII (4.51)	3.95	1.96	2.96	0.23	0.28	0.45
IX (5.28)	3.59	2.24	2.92	0.48	0.68	0.40
X (5.64)	3.83	2.06	2.94	0.47	0.60	0.18
XI (5.69)	3.93	2.17	3.05	0.47	0.68	0.29
XII (5.65)	3.52	2.21	2.86	0.28	0.38	0.07
XIII (5.60)	3.58	1.99	2.78	0.53	0.80	0.11
XIV (5.71)	3.79	1.84	2.82	0.92	0.62	0.21
XV (5.55)	3.58	2.21	2.88	0.62	0.57	0.15

Для того чтобы охарактеризовать тот предел, к которому стремится лексическая обусловленность в данном языке и его стилях, мы введем понятие «*предельная лексическая обусловленность*» (L_∞), аналогичное понятию «предельная контекстная обусловленность» (см. выше стр. 65). При этом

$$L_\infty = I^{(л)} - I_\infty^{(л)}.$$

Соотнося значения L_∞ с предельной контекстной обусловленностью, суммирующей все контекстные связи, обнаруживаем, что предельная лексическая обусловленность единицы текста при расчете по I' составляет от 23

до 33% от суммарной контекстной обусловленности (ср. таблицу 17).

Таблица 17

Лексическая и общая контекстная обусловленность (дв. ед.) во французском тексте (данные взяты из работы [69, tableau 2])

	$I_I^{(л)}$ в дв. ед.	$I'^{(л)}_\infty$ в дв. ед.	l	K'_∞ в дв. ед.	L'_∞ в дв. ед.	$K' - I'$ в дв. ед.	$\frac{L'_\infty}{K'_\infty}$ в %
Разговорная речь	3.08	2.33	0.50 (0.179)	3.26	0.74	2.52	23
Беллетристика	3.02	2.21	0.24 (0.131)	3.38	0.81	2.57	24
Деловые тексты	3.27	2.10	0.39 (0.192)	3.54	1.17	2.37	33
Язык в целом	3.15	2.25	0.36 (0.105)	3.36	0.90	2.46	27

Примечание. В скобках показаны величины среднего квадратического отклонения.

Как уже говорилось, наибольшей избыточностью и контекстной обусловленностью (соответственно наименьшей информацией) обладает деловой стиль. Большая избыточность в этой выборке является следствием высокой лексической обусловленности единиц в деловом тексте, (ср. таблицу 17).

Причины этой высокой лексической обусловленности, которая во французском деловом тексте составляет 33% суммарной контекстной обусловленности буквы в тексте, следует искать, во-первых, в использовании большого количества устойчивых словосочетаний, связанных с той или иной тематикой, во-вторых, в сравнительно ограниченном круге лексики, значительную часть которой образует терминология данной специальности, в-третьих, в логическом и строго нормализованном построении предложений.

Более низкая избыточность беллетристического стиля является результатом большей по сравнению с деловой речью неопределенностью в выборе языковых элементов. Хотя беллетристический стиль также использует строго

нормализованный синтаксис, лексические связи здесь значительно слабее: языковые штампы применяются гораздо реже, вместе с тем в этом стиле используется большое количество неожиданных словосоединений (метафоры и другие «фигуры стиля»), а круг лексики здесь гораздо шире, чем имеет место в деловой речи.

* * *

Используя выражение (47), Г. П. Богуславская (8) попыталась определить рост лексических и грамматических связей в английском тексте. Для этого были определены средние значения энтропии не только для двух первых, но и для двух последних букв слов. Затем были отдельно получены значения коэффициента a и предельного значения информации для начал и концов слов. При этом оказалось, что значение коэффициента a, характеризующего скорость нарастания связей в тексте, для кривой, проходящей по началам слов, более чем в два раза меньше величины a, характеризующей нарастание связей по концам слов.

Если согласиться с тем, что ход кривой, проходящей по началам слов, отражает нарастание лексических связей, а кривая, проходящая по их концам, говорит об усилении грамматических связей, опирающихся в основном на согласование и управление, то можно утверждать, что лексические связи текста растут медленнее, чем грамматические связи между словами.

Разумеется, следует рассматривать результаты этого эксперимента как имеющие предварительный характер. Дальнейшее исследование распределения информации в тексте должно учитывать характеристику функциональных стилей. Эти исследования следует проводить на основе достоверных совокупностей текстов, однотипных как по длине, так и по стилевой принадлежности.

§ 22. УГАДЫВАНИЕ БУКВ ТЕКСТА НА ОСНОВЕ ИНОГО ЯЗЫКОВОГО КОДА ИЛИ НЕПОЛНОГО ВЛАДЕНИЯ ТЕМ ЯЗЫКОМ, НА КОТОРОМ НАПИСАН ТЕКСТ

До сих пор угадывание букв текста осуществлялось при условии, что информант является носителем того языка, на котором написан текст. Однако ничто не мешает нам поставить такой эксперимент, в котором угадчик либо использует другой языковой код (см. выше, стр. 7), либо опирается на неполное владение тем языком, на котором написан текст.

В первом случае речь идет об угадывании текста таким информантом, который не знает языка текста и угадывает буквы текста исходя из кода другого языка. Этот язык может быть близкородственным, далекородственным или вообще неродственным языку, на котором написан контрольный текст. Совершенно очевидно, что угадывание в этом случае будет давать иные результаты, чем угадывание на родном языке. Энтропия опыта, осуществленного с испытуемым, опирающимся на неродственный язык, будет, вероятно, значительно превышать энтропию, получаемую при проведении опыта с носителем языка. Меньше расхождений дает опыт с носителем далекородственного языка, еще большее схождение покажет, очевидно, опыт с носителем близкородственного языка.

Таким образом, сравнивая количественные оценки энтропии опытов по угадыванию букв текста носителем языка с аналогичными результатами, получаемыми от носителей близкородственных, далекородственных и неродственных языков, мы получаем возможность количественно оценить степень близости различных языков в комплексе их грамматики, фразеологии, лексики, фонетики и орфо-

Таблица 18

Результаты коллективного угадывания французских и испанских текстов студентами соответствующих отделений факультетов иностранных языков и носителями языка по значениям $I_\infty = H_\infty$

Язык	I курс	II курс	III курс	IV курс	V курс и выше педагогические курсы	Носители языка
Французский (данные М. А. Зельцман)	1.74	1.39	1.35	1.32	1.28	1.09
Испанский (данные В. Я. Шабес)	1.81	1.32	1.15	0.97	—	1.05

Примечание. Для французского языка $I'_\infty = \frac{I_\infty + I_\infty}{2}$.

графии. Исходя из этих общих соображений, был осуществлен эксперимент по количественной оценке взаимного родства романских [20] и славянских (русский, белорусский, болгарский) языков.

Во втором случае речь идет об угадывании текста лицом, недостаточно владеющим иностранным языком. Чем хуже владеет угадчик языком, на котором написан контрольный текст, тем выше будут количественные оценки энтропии опыта. Сравнение этих оценок с оценками энтропии текста, получаемыми от носителя языка, могут служить мерой степени владения языком.

Все эти соображения легли в основу эксперимента, поставленного в Ленинградском пединституте им. А. И. Герцена и в Минском институте иностранных языков. Цель этого эксперимента, опирающегося на коллективное угадывание текста, состоит в том, чтобы оценить степень владения студентами изучаемого ими в качестве основной специальности иностранного языка.

При этом выясняется, что пассивное владение иностранным языком у студентов выпускного курса приближается, как показывают данные таблицы 18, к аналогичному владению его носителями (этот вид владения языка оценивался с помощью коллективного эксперимента, осуществлявшегося без ограничения времени угадывания). Выяснилось также, что превышение величин энтропии в только что описанных экспериментах по сравнению с тем уровнем энтропии, который дает обычный эксперимент, образуется за счет того, что эксперимент с иностранцем дает заметно более высокие значения энтропии для букв, находящихся в середине и конце слова.

Глава III.

ИНФОРМАЦИОННЫЕ ХАРАКТЕРИСТИКИ СЛОВА

§ 23. МЕТОДИКА ВЫВЕДЕНИЯ ИНФОРМАЦИОННЫХ СХЕМ СЛОВА

При оценке общего количества и размещения информации в слове или в словоформе (в дальнейшем мы будем для обозначения обоих этих понятий пользоваться термином «слово») методика эксперимента и расчетов остается в целом той же, что и при исследовании текста. Здесь снова применяется индивидуальное или коллективное побуквенное угадывание, осуществляющееся либо относительно слов, взятых вне контекста (эти слова мы будем называть внеконтекстными), либо при условии, что испытуемому известен контекст, предшествующий угадываемому слову (такие слова мы будем называть контекстными). В выборку внеконтекстных слов включаются обычно слова, стоящие в начале тех связных текстов, которые были подвергнуты угадыванию. Наборы контекстных слов пополняются обычно словами, стоящими на пятом, шестом и т. д. местах тех же текстов.

В тех случаях, когда необходимо получить информационную схему слова на буквенном уровне, для выборки слов определенной длины строятся обобщающие матрицы, в которых все результаты угадывания для первых букв слова попадают в первый столбец, для вторых букв — во второй и т. д. (ср. таблицу 2). Каждый столбец обрабатывается по формулам (17), (22) или (26).

В тех случаях, когда нужно определить распределение информации в слове на слоговом и морфемном уровнях, результаты угадывания группируются таким образом, чтобы получить распределения частот попыток для угадывания первой и последней букв слога или морфемы. Эта процедура, аналогичная перегруппировке экспериментального

материала при построении лексической схемы текста (см. § 21 и таблицу 17), дает возможность не только определить общее количество информации, приходящееся на слог или морфему, но позволяет также численно оценить информацию на слоговых и морфемных границах, выясняя при этом особенности слогового и морфемного членения слова.

При построении побуквенных распределений пробел считается последней буквой слова. В слоговых схемах

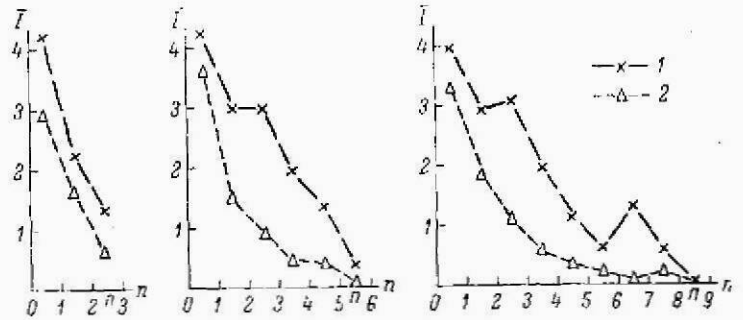

Рис. 5. Полигоны побуквенного распределения информации $\bar{I}$ в русских словах длиной в 3, 6 и 9 букв.

Условные обозначения: *1* — внетекстовое слово; *2* — текстовое слово.

пробел не учитывается (как известно, пробел не входит в состав фонетического слога). В морфемных схемах учитываются лишь те пробелы, которые выступают в функции нулевых морфем.

Следует иметь в виду, что исследование информационного членения слова может быть успешно осуществлено лишь при условии, что предшествующее расчетам выделение слогов и морфем внутри слова проводится на основе строгой и последовательной лингвистической процедуры [ср. 5].

Побуквенное, слоговое и морфемное распределения информации в словах определенной длины относятся лишь к частным случаям, не дают полного представления об общем распределении информации и мало что говорят об общих принципах слогового и морфемного членения слова. Поэтому возникает необходимость построить обобщающую информационную схему слова на всех трех уровнях.

Учитывая эмпирические вероятности появления слов определенной длины, можно свести их побуквенные схемы в обобщающую схему распределения информации в слове. Используются два вида такого обобщения.

Во-первых, может быть построена схема, соответствующая средней длине слова в данном языке. При построении

Таблица 19

Распределение информации по $\bar{I}$ в обобщающей схеме русского текстового слова, соответствующей средней длине русского слова

№№ п-п	λ	P_λ	1-я буква $\bar{I}$	P_λ^1	2-я буква $\bar{I}$	Средние буквы (1,37 буквы) $\bar{I}$	P_λ^{cp}	Предпоследняя буква $\bar{I}$	$P_\lambda^{пр}$	Последняя буква $\bar{I}$	Пробел $\bar{I}$
1	2	0.11	3.10	—	—		—	—	—		0.84
2	3	0.11	2.98	—	—		—	—	0.12	1.65	0.62
3	4	0.10	3.03	0.13	1.83		—	—	0.11	1.04	0.31
4	5	0.08	3.52	0.10	2.05		0.12	0.61	0.10	0.47	0.05
5	6	0.13	3.66	0.16	1.46		0.19	0.47	0.14	0.39	0.06
6	7	0.11	3.61	0.14	1.88		0.16	0.30	0.12	0.25	0.05
7	8	0.11	3.49	0.14	2.33		0.16	0.29	0.12	0.20	0.00
8	9	0.09	3.38	0.12	1.80		0.13	0.07	0.10	0.20	0.00
9	10	0.05	2.82	0.07	1.85		0.08	0.25	0.06	0.11	0.00
10	11	0.04	3.59	0.05	2.25		0.06	0.24	0.05	0.00	0.00
11	12 и более	0.07	3.50	0.09	1.90		0.10	0.26	0.08	0.15	0.05
12	$\bar{I}_n^{(c)}$		3.45		1.90	1.75		0.33		0.51	0.21

этой схемы буквы, или, точнее, информации, падающие на эти буквы, группируются так, как это делалось при расчете лексической схемы текста (см. § 20). Информации ($\bar{I}_n^{(c)}$), приходящиеся на первую, вторую, предпоследнюю, последнюю буквы и пробел в обобщающей схеме, получаются как суммы взвешенных информаций, приходящихся на соответствующие буквы в частных схемах. Иными словами,

$$\bar{I}_n^{(c)} = \Sigma P_\lambda \bar{I}_n^{(c)}, \qquad (52)$$

где P_λ — эмпирическая вероятность слова определенной длины (λ), $\bar{I}_n^{(c)}$ — информация, приходящаяся на n-ю

букву в частной схеме. Количество информации, содержащееся в буквах, которые находятся в интервале между второй и предпоследней буквами усредненной схемы, получается как среднее арифметическое первой, второй, предпоследней и последней букв слова, умноженное на количество букв в этом интервале. Расчет обобщающей схемы русского текстового слова по $\bar{I}$ дан в таблице 19.

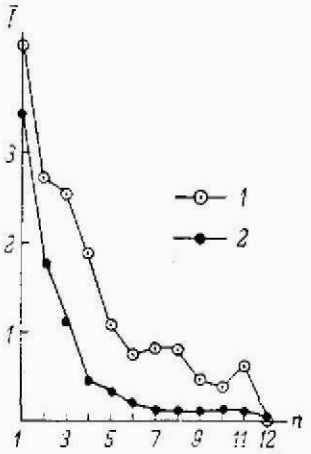

Рис. 6. Обобщающая схема русского слова без ограничения длины, построенная по $\bar{I}$.

Условные обозначения: 1 — внетекстовое слово; 2 — текстовое слово.

Во-вторых, может быть построена обобщающая схема, на которую не накладываются ограничения, связанные со средней длиной слова в текстах данного языка. Поскольку предел длины слова ничем не ограничен, и длина обобщающей схемы в этом случае стремится к бесконечности. Однако практически оказывается, что очень длинные слова встречаются в языке сравнительно редко. Так, например, в русском языке слова длиной более одиннадцати букв, а во французском и английском — более десяти букв составляют в среднем не более 5—7% общего числа слов. Поэтому целесообразно ограничить рассматриваемую схему некоторым пределом. В русском языке, например, эта схема доведена до двенадцатой буквы (11 букв+пробел) [см. 24, стр. 113]. Расчет этой схемы для русского внетекстового и текстового слова по $\bar{I}$ и I дан в таблице 21, ее чертеж относительно $\bar{I}$ см. на рис. 6.

Аналогичным путем строятся обобщающие схемы слов на *слоговом* и *морфемном* уровнях. Ниже в качестве образца приводятся таблицы 22, 23, содержащие данные для *построения обобщающей слоговой и морфемной* схем русского внетекстового слова. Чертежи этих схем см. на рис. 7, ср. также [24, стр. 114—115].

При оценке достоверности частных и общих информационных схем слова используются, как уже говорилось (см. § 17), контрольные угадывания на других испытуе-

Таблица

Распределение информации в русских словоформах различной длины, взятых в тексте (в дв. ед.)

№ п/п	Длина словоформы в буквах (λ), включая пробел	Вероятность словоформы данной длины $P_{(\lambda)}$	Информация на букву																				Общее количество информации, приходящееся на словоформу определенной длины		Среднее количество информации, приходящееся на букву		Общее количество грамматической информации, приходящееся на словоформу		Среднее количество грамматической информации приходящееся на букву	
			1-я		2-я		3-я		4-я		5-я		6-я		7-я		8-я		9-я		10-я		11-я		12-я							
			I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1	2	0.11	3.10	3.10	0.38	0.84	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3.48	3.94	1.74	1.97	0.00	0.00	0.00	0.00
2	3	0.11	2.83	2.98	0.95	1.65	0.28	0.62	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	4.06	5.25	1.35	1.75	0.24	0.35	0.08	0.11
3	4	0.10	3.50	3.63	1.00	1.83	0.57	1.04	0.14	0.31	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	5.21	6.81	1.30	1.70	0.59	1.20	0.15	0.30
4	5	0.08	3.26	3.52	1.25	2.05	0.24	0.61	0.20	0.47	0.03	0.05	—	—	—	—	—	—	—	—	—	—	—	—	—	—	5.08	6.70	1.01	1.32	0.61	1.05	0.12	0.21
5	6	0.13	2.94	3.66	0.82	1.46	0.43	0.81	0.24	0.47	0.20	0.39	0.02	0.06	—	—	—	—	—	—	—	—	—	—	—	—	4.65	6.85	0.78	1.14	0.36	0.72	0.06	0.12
6	7	0.11	2.57	3.61	1.14	1.88	0.55	1.08	0.21	0.44	0.13	0.30	0.12	0.25	0.02	0.05	—	—	—	—	—	—	—	—	—	—	4.74	7.61	0.68	1.09	0.43	0.75	0.06	0.11
7	8	0.11	2.90	3.49	1.46	2.33	0.77	1.43	0.17	0.38	0.08	0.17	0.14	0.29	0.10	0.20	0.00	0.00	—	—	—	—	—	—	—	—	5.62	8.29	0.70	1.04	0.40	0.67	0.05	0.08
8	9	0.09	2.79	3.38	1.02	1.80	0.56	1.06	0.25	0.52	0.16	0.30	0.10	0.21	0.02	0.07	0.07	0.20	0.00	0.00	—	—	—	—	—	—	4.97	7.54	0.55	0.84	0.37	0.72	0.04	0.08
9	10	0.05	3.20	3.82	1.04	1.85	0.93	1.66	0.18	0.42	0.19	0.41	0.04	0.11	0.04	0.10	0.12	0.25	0.04	0.11	0.00	0.00	—	—	—	—	5.78	8.73	0.58	0.87	0.42	0.49	0.04	0.05
10	11	0.04	2.94	3.59	1.43	2.25	1.63	2.24	0.51	0.88	0.10	0.24	0.10	0.22	0.00	0.00	0.12	0.21	0.12	0.24	0.00	0.00	0.00	0.00	—	—	6.95	9.82	0.63	0.90	0.30	0.56	0.03	0.05
11	12	0.07	2.88	3.50	1.12	1.90	0.91	1.58	0.44	0.84	0.28	0.56	0.24	0.40	0.14	0.27	0.00	0.00	0.04	0.10	0.12	0.26	0.07	0.15	0.04	0.05	6.28	9.61	0.52	0.80	0.60	0.98	0.05	0.08
12	Среднее количество информации, приходящееся на букву в обобщающей схеме слоформы		2.98	3.45	1.02	1.75	0.59	1.11	0.22	0.46	0.14	0.30	0.11	0.20	0.06	0.12	0.05	0.11	0.04	0.09	0.06	0.11	0.04	0.10	0.04	0.05	5.99	8.15	0.94	1.28	0.38	0.70	0.06	0.11

Таблица 20

Распределение информации в русских словоформах различной длины, взятых вне текста (в дв. ед.)

№ п./п.	Длина словоформы в буквах (λ), включая пробел	Вероятность словоформы данной длины $P(\lambda)$	Информация на букву																				Общее количество информации, приходящееся на словоформу определенной длины		Среднее количество информации, приходящееся на одну букву словоформы		Общее количество грамматической информации, приходящееся на словоформу		Среднее количество грамматической информации, приходящееся на букву	
			1-я	2-я		3-я		4-я		5-я		6-я		7-я		8-я		9-я		10-я		11-я		12-я								
			$I=I$	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
			1—2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1	2	0.11	4.22	0.55	1.09	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	4.77	5.31	2.38	2.66	0.00	0.00	0.00	0.00
2	3	0.11	4.22	1.53	2.25	0.81	1.34	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	6.35	7.81	2.19	2.60	0.20	0.38	0.07	0.13
3	4	0.10	4.22	2.13	2.80	1.26	1.95	0.29	0.58	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	7.90	9.55	1.98	2.39	1.85	3.10	0.46	0.78
4	5	0.08	4.22	2.60	3.26	1.75	2.56	1.43	2.21	0.00	0.00	—	—	—	—	—	—	—	—	—	—	—	—	—	—	10.00	12.27	2.00	2.45	2.25	3.29	0.45	0.66
5	6	0.13	4.22	2.40	2.98	2.27	2.95	1.51	1.89	0.85	1.33	0.16	0.31	—	—	—	—	—	—	—	—	—	—	—	—	11.41	13.68	1.90	2.28	2.30	3.52	0.38	0.59
6	7	0.11	4.22	3.23	3.26	1.66	2.31	1.15	1.71	1.02	1.36	0.52	0.98	0.23	0.43	—	—	—	—	—	—	—	—	—	—	12.03	14.27	1.72	2.04	1.43	2.27	0.20	0.32
7	8	0.11	4.22	2.83	3.14	3.18	3.28	1.53	2.19	0.73	0.92	0.73	1.10	0.77	1.28	0.02	0.26	—	—	—	—	—	—	—	—	14.01	16.37	1.75	2.05	1.27	2.05	0.16	0.26
8	9	0.09	4.22	2.12	2.88	2.88	3.08	1.37	1.92	0.88	1.13	0.34	0.61	0.62	1.35	0.30	0.66	0.00	0.00	—	—	—	—	—	—	12.93	15.85	1.44	1.76	1.70	2.87	0.19	0.32
9	10	0.05	4.22	2.30	2.89	2.66	3.19	2.21	2.60	0.93	1.43	0.20	0.58	0.32	0.51	0.96	1.42	0.96	1.32	0.00	0.00	—	—	—	—	11.76	18.16	1.48	1.82	3.50	4.78	0.35	0.48
10	11	0.04	4.22	2.51	2.95	2.61	3.08	1.38	2.45	0.69	1.18	0.74	1.21	0.46	0.72	0.40	0.64	0.29	0.56	0.58	0.89	0.32	0.64	—	—	13.84	18.14	1.27	1.65	1.81	2.88	0.17	0.26
11	12 и более	0.07	4.22	2.60	2.99	2.43	2.62	1.65	2.14	0.75	1.20	0.66	0.69	0.18	0.33	0.38	0.62	0.32	0.53	0.20	0.40	0.54	0.71	0.00	0.00	13.93	16.45	1.16	1.37	1.41	2.12	0.12	0.18
12	Среднее количество информации, приходящееся на букву в обобщающей схеме словоформы		4.22	2.20	2.73	2.01	2.52	1.31	1.87	0.74	1.06	0.46	0.75	0.45	0.81	0.37	0.81	0.33	0.50	0.23	0.40	0.46	0.65	0.00	0.00	14.59	13.50	1.82	2.12	1.46	2.30	0.23	0.36

Таблица 22

Слоговая схема внетекстовой русской словоформы по $\bar{I}$

№ п/п	Длина слова в слогах	Вероятность появления слова данной длины	Информация в дв. ед. по слогам											
			I		II		III		IV		V		VI	
			первая буква	последняя буква	первая буква	последняя буква	первая буква	последняя буква	первая буква	последняя буква	первая буква	последняя буква	первая буква	последняя буква
1	1	0.339	4.22	2.25										
2	2	0.303	4.22	2.85	2.60	1.35								
3	3	0.214	4.22	3.00	3.26	1.09	0.82	1.05						
4	4	0.097	4.22	2.16	3.12	1.66	1.58	1.12	0.88	1.07				
5	5	0.035	4.22	2.15	2.57	1.23	0.74	0.62	0.49	0.35	0.71	0.57		
6	6 и более	0.012	4.22	2.56	2.32	1.48	1.24	0.38	0.60	0.18	0.69	1.23	0.75	0.40
7	Обобщающая слоговая схема словоформы		4.22	2.59	2.89	1.31	1.03	1.08	0.77	0.82	0.74	0.74	0.75	0.40
8	Средняя длина слога в буквах		2.30		2.20		2.39		2.30		1.95		1.55	

Таблица 23

Морфемная схема внетекстовой русской словоформы по $\bar{I}$

№№ пп	Длина слова в морфемах	Вероятность появления слов данной длины	Информация в дв. ед. морфемы I		II		III		IV		V	
			первая буква	последняя буква	первая буква	последняя буква	первая буква	последняя буква	первая буква	последняя буква	первая буква	последняя буква
1	1	0.260	4.22	1.87								
2	2	0.417	4.22	2.28	2.24	1.05						
3	3	0.165	4.22	1.80	2.30	1.96	1.72	0.73				
4	4	0.118	4.22	2.02	2.93	0.99	1.31	0.00	1.18	0.56		
5	5 и более	0.040	4.22	2.17	2.84	0.39	1.49	0.00	1.77	0.80	1.11	0.79
6	Обобщающая морфемная схема в словоформах		4.22	2.03	2.39	1.19	1.54	0.35	1.39	0.63	1.11	0.79
7	Средняя длина морфемы в буквах		3.24		2.07		1.70		1.70		1.44	

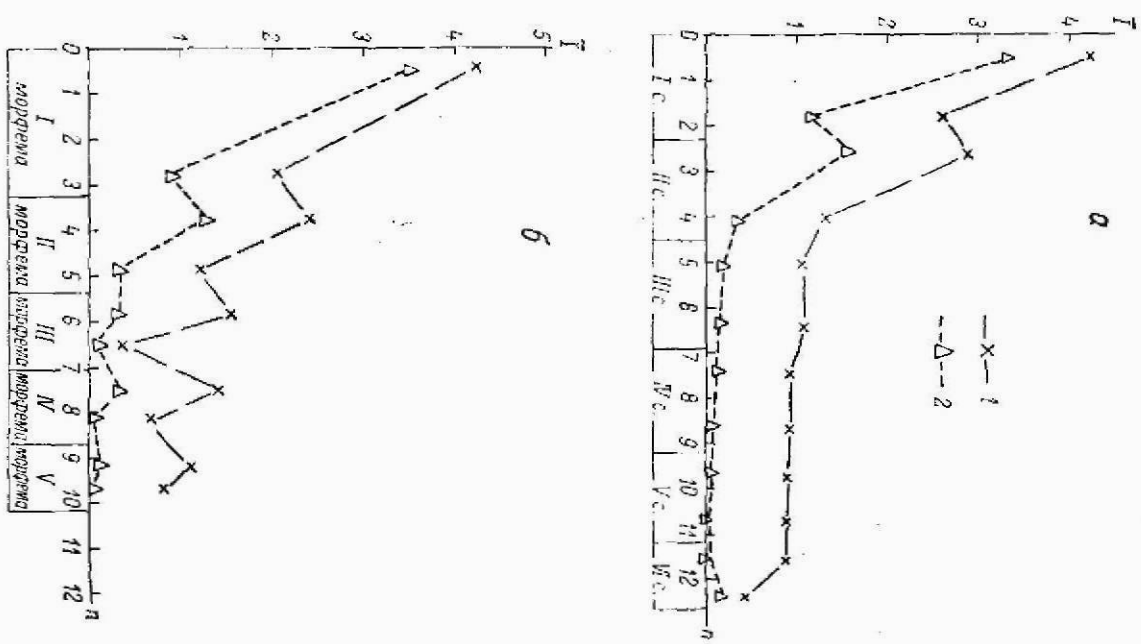

Рис. 7. Полигоны распределения информации I в обобщающих слоговых (*а*) и морфемных (*б*) схемах русского слова.

Условные обозначения: *1* — полигоны внетекстового слова; *2* — полигоны текстового слова.

мых, коллективное угадывание, а также сопоставление с данными, полученными путем расчета вероятностных спектров слов (частотных словарей).

§ 24. БУКВЕННОЕ, СЛОГОВОЕ И МОРФЕМНОЕ СТРОЕНИЕ СЛОВА

Распределение информации в коротких (до четырех букв), средних (от пяти до семи букв) словах, с одной стороны, и длинных словах (от восьми букв и выше) — с другой, имеет разный характер. Короткие и средние слова, как правило, дают монотонное убывание информации от начала слова к его концу. Убывание это обычно происходит гладко, и полигоны (графические схемы) этих слов имеют компактный вид (ср. рис. 5). Такое распределение связано с тем, что среди коротких и длинных слов достаточно часто употребляются неизменяемые формы (ср. русск. *о*, *да*, *еще*, *если*, *очень*, *только*; англ. *a*, *of*, *and*, *for*, *over*, *alone*, *almost*; фр. *à*, *on*, *oui*, *très*, *aussi*, *encore*; рум. *o*, *de*, *mai*, *dacă*, *afară*, *pentru*), которые в суммирующей схеме сглаживают максимумы информации на конечных аффиксах изменяемых слов.

У длинных слов (в первую очередь у внетекстовых) полигоны принимают постепенно так называемую U-образную форму: максимумы информации сосредоточены здесь в начале и в конце словоформы, буквы же, находящиеся в средней части словоформы, несут наименьшую информацию (ср. рис. 5). U-образный характер распределения информации в длинных словоформах отражает определенные вероятностно-статистические и лингвистические закономерности. Присутствие максимума информации в начале слова отражает тот факт, что начальные буквы письменного слова содержат наибольшее количество неопределенности в выборах. Отсюда следует, что указанные буквы несут наибольшее количество информации. Это с интуитивной точки зрения правдоподобное предположение было проверено и количественно оценено с помощью экспериментов по угадыванию на материале русского, английского, французского, испанского и румынского (молдавского) языков [см. 17, стр. 104—105; 25, стр. 125; 49; 70]. Оно было также подтверждено путем проведения коллективных экспериментов по восстановлению начальных и конечных букв внетекстовых и текстовых слов.

Появление максимума информации в конце слова обусловлено, очевидно, морфологической структурой слова в исследованных языках, в которых бóльшая часть грамматической информации концентрируется в присоединяемых к концу слова аффиксах (окончаниях и суффиксах).

Тот факт, что морфологическая структура изменяемого слова обнаруживается яснее всего в информационных схемах длинных словоформ, объясняется тем, что в этих схемах обобщаются почти исключительно словоформы, имеющие приставки и конечные аффиксы.

Буквенные распределения, в которых обобщаются формы слов разных структур (изменяемые и неизменяемые, односложные и многосложные), дают весьма приближенную и грубую схему распределения информации. Для того чтобы глубже проникнуть в информационное строение слова, были проанализированы распределения информации в русских, английских, французских и румынских словоформах на слоговом и морфемном уровнях. Полные количественные данные по частным и общим распределениям, а также их чертежи приводится в работах [24, стр. 114—116; 31, стр. 155—159; 68а].

Построение слоговых и морфемных схем таково, что позволяет наблюдать межслоговые «швы» и границы между морфемами, образующими слово.

Схемы слогового распределения информации показывают, что во всех трех языках слоговое деление слова обнаруживается лишь на границе первого и второго слова или, иначе говоря, в начале хода кривой распределения (ср. в этом смысле чертежи обобщающих слоговых схем, приведенные на рис. 7, *а*). По мере продвижения кривой вправо слоговые границы все более затушевываются, а начиная с 4-го слога полностью исчезают.

Иную картину дает морфемное построение слова. Морфемные схемы внетекстовых слов во всех трех языках характеризуются последовательным чередованием максимумов и минимумов информации. Максимумы информации падают на первую, а минимумы на последнюю букву морфемы. Иными словами, на всем протяжении внетекстового письменного слова оказываются четко обозначенными границы между морфемами. Эти границы совпадают с водоразделом между последней буквой предшествующей морфемы (минимум информации) и первой буквой после-

дующей морфемы (максимум информации). Если обратиться к текстовым словам, то окажется, что они также обнаруживают морфемное членение, хотя оно и выражено здесь значительно слабее (ср. рис. 7, *б*).

Итак, можно утверждать, что слову в русском, английском, французском и румынском языках присуща отчетливо выраженная морфемная структура, которая подавляет с точки зрения статистической информации слоговое членение слова.[17] Так обстоит, очевидно, дело в тех языках, в которых слог не совпадает с морфемой.

Если только что приведенные результаты подтвердятся на материале других языков, то прояснятся некоторые аспекты семиотической природы языка и ее функционирования в речи.

Язык реализуется в речи. Речь же представляет собой сложный марковский процесс линейного следования различных элементов языка — как фигур, так и знаков. В связи с этим возникает вопрос: какие закономерности лежат в основе вероятностно-статистической (теоретико-информационной) структуры речи? Вероятностно-статистические закономерности, характеризующие сочетаемость фигур, или закономерности сочетаемости знаков [ср. 58, стр. 305]? Исследованный материал показывает, что вероятностно-статистические связи и ограничения, характеризующие сочетаемость фигур (букв, фонем, слогов), действуют на коротких начальных участках текста (в нашем случае — на первых буквах словоформы), не превышающих длины кратчайшего знака (чаще всего морфемы), в который входят данные фигуры. Как только следующие друг за другом по синтагматической оси фигуры сформируют знак, на сцену выступают вероятностно-статистические закономерности сочетаемости знаков. Эти закономерности накладываются на вероятностные спектры сочетаемости последующих фигур, отбирая из этих спектров лишь такие сочетания, которые соответствуют правилам сочетаемости знаков, образованных из этих фигур.

Отсюда следует, что широко проводимые сейчас работы по статистике и комбинаторике букв, буквосочетаний и слогов дают слишком мало сведений о структуре конкретного языка (или функционального стиля). Результаты таких работ успешно могут быть использованы лишь в качестве вспомогательного аппарата для изучения языка на уровне знаков.

[17] В отношении английского языка это предположение находит подтверждение в исследовании Д. Карсона [49, стр. 50], посвященном вероятностно-статистическим ограничениям внутри английской письменной словоформы. Оказалось, что процентное соотношение гласных и согласных (или, иначе говоря, слоговая структура словоформы) влияет на ход кривой этих вероятностно-статистических ограничений лишь относительно первых двух букв. Вместе с тем ни русский, ни английский материал не подтверждает неоднократно высказывавшихся предположений о том, что максимумы или минимумы информации (энтропии) могут рассматриваться в качестве объективных показателей границ слога [22, стр. 190; 25, стр. 125]. Пики информации могут отмечать границы слога лишь в начальных участках текста (словоформы), длина которых не превышает длины первой морфемы.

§ 25. КОНТЕКСТ И ИНФОРМАЦИОННАЯ СТРУКТУРА СЛОВА

Понятие контекстной обусловленности, примененное в § 19 при исследовании текста, может быть использовано и в ходе информационного анализа структуры слова.

Снова предположив, что в языке не существует комбинаторно-статистических ограничений и что все буквы и другие единицы обладают неограниченными возможностями сочетаемости, мы можем определить то оптимальное количество информации, которое несет слово длиной в λ букв. Эта информация равна

$$I_0^{(\lambda)} = \lambda I_0.$$

В действительности слово несет гораздо меньше информации, что является, как уже указывалось, результатом разного вида статистических ограничений. Разность между оптимально возможным и реальным количеством информации, содержащимся в слове ($I^{(\lambda)}$), мы будем называть общей контекстной обусловленностью слова ($K^{(\lambda)}$). При этом

$$K^{(\lambda)} = I_0^{(\lambda)} - I^{(\lambda)}. \qquad (53)$$

Общая контекстная обусловленность слова является функцией разных лингвостатистических ограничений. У внетекстового слова величина контекстной обусловлен-

ности характеризует сумму статистических ограничений, связанных с сочетаемостью букв, слогов и морфем. Эту величину мы будем называть *внутрисловной обусловленностью* ($K^{(вс)}$). У текстовой словоформы, содержащей $I^{(ст)}$ дв. ед. информации, контекстная обусловленность ($K^{(ст)}$) представляет собой сумму внутрисловной обусловленности и той обусловленности, которую вносят лексико-грамматические связи между словами. Эту последнюю мы будем называть *лексико-грамматической обусловленностью* ($K^{(лг)}$).

Таблица 24

Доля ограничений, вносимых лексико-грамматическими связями между словами в тексте, в общей контекстной обусловленности слова

Язык	%
Русский (по I)	22.5
Английский (по I') . .	35.0
Французский (по I') . .	23.7
Румынский (молдавский) предварительные данные по I' . .	27.1

Нетрудно показать, что

$$K^{(лг)} = I^{(с)} - I^{(ст)}. \quad (54)$$

Численные оценки контекстных обусловленностей слова в русском, английском, французском и румынском (молдавском) языках уже приводились в ряде статей [см. 8, 24, стр. 119; 31, стр. 167; 68а].

Для языкознания наибольший интерес представляют не столько абсолютные величины контекстных обусловленностей, сколько их отношения в рамках того или иного языка. В частности, важно определить, какую долю в общей контекстной обусловленности составляют статистические ограничения, вносимые лексико-грамматическими связями между словами в тексте. Доля этих ограничений в процентах может быть определена из следующего выражения:

$$W = \frac{I^{(с)} - I^{(ст)}}{I_0^{(с)} - I^{(ст)}} 100 = \frac{K^{(лг)}}{K^{(с)}} 100.$$

Значения W для четырех европейских языков показаны в таблице 24.

Таковы общие количественные оценки функциональной отдачи лексико-грамматического контекста. Теперь попытаемся выяснить некоторые детали воздействия указанного контекста на информационную структуру слова.

Первое, что бросается в глаза при сопоставлении текстовых (нижние кривые на рис. 6 и 7) и внетекстовых (верхние кривые на тех же рисунках) графиков, это не только значительное уменьшение общего количества информации у текстового слова, но и изменение общей конфигурации распределения. Обобщающие полигоны морфемного распределения информации (см. рис. 7) показывают, что лексико-грамматический контекст, срезая максимумы информации на первых буквах морфем, расположенных в середине и конце слов, в значительной степени сглаживает границы между морфемами.

Сходную картину дают полигоны буквенных распределений (см. рис. 5 и 6). U-образные распределения информации, характерные для длинных русских и французских слов, употребленных вне контекста, сменяются в контексте J-образной конфигурацией (аналогичную картину дают английский, французский и румынский языки, см. [5, стр. 104—108; 24, стр. 114—116; 68а]. Это происходит потому, что лексико-грамматический контекст срезает максимумы информации, расположенные в конце слова. А эти максимумы, как уже указывалось, концентрируют в себе основную часть грамматической информации, содержащейся в слове.

§ 26. ИЗМЕРЕНИЕ ГРАММАТИЧЕСКОЙ ИНФОРМАЦИИ

Следует сразу подчеркнуть, что речь пойдет о грамматической информации, заключенной в префиксах, суффиксах, внутренних флексиях и окончаниях слова. Что же касается информации, передаваемой словообразовательными аффиксами и средствами аналитической морфологии, то она в расчет приниматься не будет.

Грамматическая информация, содержащаяся в слове (точнее, в схеме слова определенной длины) определяется как сумма грамматических информаций, падающих на каждую буквенную позицию в этой схеме.

Для определения грамматической информации, которую несет данная буквенная позиция, обобщенные в этой позиции буквы экспериментальных слов делятся на два разряда. В первый включаются те буквы, которые либо

входят в состав грамматического аффикса (окончание, внутренняя флексия, приставка), либо хотя и не составляют грамматической части слова, но имеют альтернативой другую букву, входящую в грамматический аффикс данного или другого конкретного слова. Второй разряд составляют буквы, не входящие в грамматический аффикс и не имеющие грамматических альтернатив.

Буквы, входящие в первый разряд, группируются по количеству попыток, понадобившихся для их отгадывания. Все буквы второго разряда относятся к группе достоверных продолжений (угадывания с «нулевой» попытки) независимо от того, сколько попыток понадобилось, чтобы угадать каждую букву. Это делается из тех соображений, что «неграмматические» буквы, равно как и буквы первого разряда, угаданные с «нулевой» попытки, не несут грамматической информации. Полученное этим путем распределение вероятностей рассчитывается с помощью формул (17), (22), (26).

Данные о количестве грамматической информации, содержащейся в русских внетекстовых и текстовых словах определенной длины, в усредненных внетекстовых и текстовых словах, а также сведения о среднем количестве грамматической информации, приходящемся на букву, даны в таблицах 20 и 21. Аналогичные данные для английского и французского языков можно найти в статьях [5, 24]. Эти данные показывают, например, что, попадая в контекст, русская словоформа теряет в среднем 1.60 дв. ед. грамматической информации (считая по верхней границе). Это составляет около 69% от всей грамматической информации, содержащейся в формах внетекстового слова (для нижней границы информации эта величина составляет 67%).

Французское слово, попадая в текст, теряет 1.71 дв. ед. информации (считая по I'), что составляет около 66% грамматической информации внетекстового слова [ср. 24, стр. 120—121]. Для английского языка эти величины соответственно равны 0.605 дв. ед. и 76% [см. 5].

Исходя из всего сказанного, можно предположить, что в русском, французском и, вероятно, в английском текстах значительная часть грамматических аффиксов является избыточной. Если это так, то грамматические значения, присущие этим аффиксам, должны передаваться с помощью других средств, в первую очередь с помощью служебных слов.

В этой связи было бы интересно рассмотреть с информационно-статистической точки зрения то взаимодействие, в которое вступают в тексте служебные и знаменательные слова в целом, с одной стороны, и служебные слова, основы знаменательных слов и их флексии — с другой.

§ 27. ИНФОРМАЦИОННО-СТАТИСТИЧЕСКИЕ ОЦЕНКИ ФЛЕКТИВНОЙ И АНАЛИТИЧЕСКОЙ МОРФОЛОГИИ

То, что в ряде европейских языков служебные слова функционально эквивалентны флексиям, и наоборот, окончания знаменательных слов дублируют значения служебных слов, не подлежит сомнению. Об этом можно судить, в частности, по французской фразе типа ***nous ne le connaissons pas***, которая легко преобразуется в разговорный вариант *connaissons pas* (Laffitte. Nous retournerons cueillir les jonquilles, см. [30, стр. 110]), или по такому русскому примеру, как *завтра **я** выезжа**ю** в Москву*, который в «телеграфном» стиле выглядел бы *как завтра выезжа**ю** Москву*.

Всякий текст имеет линейный характер. Вместе с тем развертывание текста осуществляется на синтагматической оси, которая имеет либо временну́ю (устная речь), либо пространственную (письменная речь) направленность. Поскольку служебные слова в рассматриваемых языках предшествуют на этой оси функционально эквивалентным им аффиксам знаменательных слов, нетрудно догадаться, что первые несут информационную нагрузку (в данном случае речь идет как о статистической, так и о семантической информации), в то время как вторые оказываются избыточными.

Не было бы необходимости останавливаться на этих общеизвестных и тривиальных положениях, если бы в большинстве европейских языков не обнаруживались отчетливые тенденции к развитию аналитизма и к ослаблению флексий. Иными словами, аналитическая техника передачи грамматических значений оказывается как бы более *удобной*, чем флективный способ их выражения. В этой ситуации интересно было бы рассмотреть информационно-статистический механизм *отношений*, существующих между служебным и знаменательным словом, с одной стороны, и основой знаменательного слова и его окончанием, с другой.

Будем считать мерой связи данного лингвистического знака с предшествующим контекстом величину, обратную сумме информаций, падающих на первые две буквы воплощающей его цепочки

$$C = \frac{1}{I_1 + I_2}.$$

Используя величины I для русского языка и I' для французского, получаем оценки C для служебных и знаменательных слов, а также для флексий знаменательных слов в тексте (см. таблицу 25).

Таблица 25

Оценки C для служебных, знаменательных слов и флексий слов в тексте

Языки	Служебное слово	Знаменательное слово	Флексия знаменательного слова
Русский (по I) . .	0.21	0.19	0.46
Французский (по I')	0.24	0.19	0.66

Приведенные в таблице 25 данные, а также аналогичные оценки, использованные в работе [26, стр. 168], показывают, что служебное слово в значительно меньшей степени зависит от предшествующего слова, чем флексия. Невелика степень предсказуемости предшествующим контекстом и самого знаменательного слова в целом. Это дает нам право предполагать, что высокую степень обусловленности флексий следует относить не за счет всего предшествующего им контекста, а за счет лексической основы знаменательного слова. Иными словами, можно утверждать, что служебные слова благодаря своему положению на синтагматической оси обладают меньшей контекстной обусловленностью и одновременно большей статистической информацией, чем флексии.

Вполне возможно, что преимущества аналитической морфологии перед морфологией флективной обусловлены еще одной информационно-статистической особенностью служебных слов. Дело в том, что короткие слова, значительную долю которых составляют служебные слова, с точки зрения количества передаваемой ими информации

оказываются менее подверженными воздействию контекста, чем средние и длинные слова. Раньше уже было показано [см. 24, стр. 120; 31, стр. 169], что в русском языке короткие слова, попадая в контекст, теряют от 30 до 33% несомой ими информации, в то время как у длинных слов и слов средней длины контекст снимает от 47 до 58% информации (для французского языка эти величины соответственно равны 25—30 и 62—87%).

Чтобы разобраться в механизме этого явления, рассмотрим рост контекстной обусловленности в усредненной схеме текстового и внетекстового слова, а также ход этой обусловленности в связном тексте.

Как уже говорилось, ход кривой контекстной обусловленности (K_ξ) описывается зависимостью

$$K_\xi = (I_0 - I_\infty)(1 - e^{-\alpha\xi}), \qquad (46)$$

в которой I_∞ характеризует тот предел, к которому стремится информация данного типа сообщения. Следует подчеркнуть, что предельная информация связного текста, при $n \to \infty$, будет всегда больше нуля. Это и понятно. Всякий текст, будучи образован из сложных знаков (слов, словосочетаний, предложений), обладающих практически неограниченной комбинаторной способностью, имеет несколько продолжений, или, иначе говоря, всегда обладает неопределенностью выбора. Даже в тех случаях, когда данный шаг конкретного текста предусматривает единственно возможное продолжение, всегда найдутся последующие шаги, которые дадут несколько возможных продолжений. Если же рассматривать некоторую совокупность текстов, на основе которой строится усредненная схема (ср. § 19), то наряду с текстами, дающими на данном шаге достоверное продолжение, всегда обнаружатся другие тексты, которые дадут на этом шаге несколько продолжений.

Иное дело слово. Слово состоит из фигур (букв, фонем, слогов) и простых знаков (морфем), обладающих ограниченной комбинаторикой, и — что самое главное — слово выступает в тексте в виде кванта информации. Поэтому при $n \to \infty$, т. е. при бесконечном удлинении слова, информация отдельных составляющих ее фигур и знаков будет стремиться к нулю.

В связи с этим выражение, описывающее рост контекстной обусловленности внутри слова, принимает вид

$$K_{\xi}^{(c)} = I_0 - I_0 e^{-s\xi},$$

Сравнение распределений общей контекстной обусловленности в усредненной схеме текстового и внетекстового слов, проведенное на русском и французском материале [ср. 24, стр. 122—123; 25, стр. 123; 31, стр. 170—171;

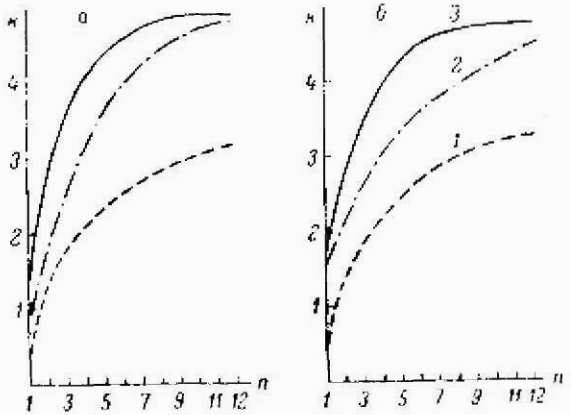

Рис. 8. Рост контекстной обусловленности в связном тексте и внутри слова в русском (*а*) и французском (*б*) языках (по оценкам *I*).

Условные обозначения: *1* — связный текст; *2* — внетекстовое слово; *3* — текстовое слово.

69, стр. 141 и 149], показало, что лексико-грамматический контекст значительно ускоряет рост контекстной обусловленности букв внутри слова. Коэффициент s текстового слова в два раза превышает по величине аналогичный коэффициент для нетекстового слова [ср. 31, стр. 167].

Особенно круто возрастает контекстная обусловленность в текстовом слове на участке от первой до четвертой букв. После четвертой буквы кривая обусловленности значительно приближается к своему пределу $K_{\infty}^{(c)}$ (для русского слова $K_{\infty}^{(c)} = I_0 = 5$ дв. ед., для французского — $K_{\infty}^{(c)} = I_0 = 4.76$ дв. ед.).

Что же касается внетекстового слова, то здесь нарастание контекстных связей происходит более плавно. Кривая обусловленности достигает своего предела лишь после двенадцатой буквы (ср. рис. 8).

Быстрое нарастание общей контекстной обусловленности букв на участке между первой и четвертой буквами — нарастание, постоянно наблюдаемое в процессе самого эксперимента, — имеет важные последствия для информационной структуры текстового слова.

Как уже говорилось, основная часть грамматической информации длинных и средних слов концентрируется на пятой, шестой и т. д. буквах. Когда слова этого типа попадают в текст, общая контекстная обусловленность их начальных букв растет настолько быстро, что несущие грамматическую информацию конечные буквы, равно как и буквы, находящиеся в центральной части слова, оказываются почти полностью предопределенными предшествующим контекстом. Поэтому они теряют значительную часть своей грамматической информации.

Иное дело короткие слова. Контекстная обусловленность при употреблении их в тексте также растет довольно быстро. Однако эта обусловленность не успевает к концу слова достичь своего предела (максимальная длина коротких слов — четыре буквы). В связи с этим все буквы короткого слова еще сохраняют свой информационный вес. Поэтому короткие слова, основную массу которых составляют служебные слова, оказываются по сравнению с длинными и средними словами, бо́льшая часть которых является знаменательными, менее подверженными воздействию контекста.

В ряде работ [1; 13; 32, стр. 58] были предприняты попытки оценить числом степень аналитичности некоторых языков. Эти оценки были получены путем сравнения общего количества слов-типов (словарных слов) и порождаемых ими словоформ. Выяснилось, что французский и английский языки дают более высокую степень аналитичности, чем русский, тексты которого, как и следовало ожидать, характеризуются значительно бо́льшим процентом флективных форм.

Однако эти оценки, хотя они и учитывают степень насыщенности текста флективными формами, еще ничего не говорят о функциональной отдаче этих форм. Эксперимент по угадыванию позволяет получить информационно-статистическую оценку такой отдачи. Эта оценка может быть, в частности, получена путем соотнесения той грамматической информации, которую несут морфологические аффиксы в усредненной схеме слова, с общим объемом

статистической информации, содержащейся в этой схеме [ср. 5, стр. 115; 24, стр. 120; 26, стр. 169]. Процентные доли этой грамматической информации по оценкам I и I' показаны в таблице 26. Сравнение долей грамматической информации приводит к неожиданным результатам.

Таблица 26

Доля информации, передаваемая флективной морфологией, в общей сумме статистической информации, падающей на слово

Язык	Внетекстовое слово, %	Текстовое слово, %
Русский (по I)	17.0	8.6
Английский (по I')	6.3	3.5
Французский (по I')	22.0	16.0
Румынский (по I')	23.0	13.0

Информационная доля флективной морфологии во французском и румынском тексте в несколько раз превосходит информационный вес английских флексий. Более того, в аналитических французском и румынском языках указанная доля флективной морфологии примерно в два раза превышает информационный вес морфологических аффиксов синтетического русского языка.

Объяснение этим соотношениям следует искать, очевидно, в следующих особенностях романских языков.

1. Глагольные парадигмы французского и румынского языков характеризуются значительным количеством флективных графических форм (ср. французские Conditionnel, Présent, Passé simple, Imparfait, Imparfait du Subjonctif), превосходя в этом отношении не только английский, но и русский язык. Употребление французских глагольных окончаний не всегда предопределено формой существительного или местоимения, ср. фр. *je (tu) chantais*, *nous chantons* (*chantions*); рум. *eu cînt (cîntam, cîntai)*.

2. Романские служебные слова, играющие большую роль в передаче грамматической информации, часто сами имеют флексии (ср. формы разных видов артикля: фр. *le, l', la, les*; *un, une*; рум. *un, unui, unei*. Что же касается русских и английских служебных слов, то они чаще всего оказываются неизменяемыми.

ЗАКЛЮЧЕНИЕ

Описанные результаты эксперимента и связанные с ними рассуждения не претендуют, разумеется, на то, чтобы их рассматривать в качестве сколько-нибудь законченной теории. Скорее это один из шагов на пути проб и ошибок, ведущих к познанию тайн функционирования языка в речи.

Попробуем оценить этот шаг с точки зрения общих вопросов методики исследования, проблем коммуникации и некоторых лингвистических задач.

Своеобразие процесса речевого общения состоит в том, что он представляет целенаправленную деятельность с постоянно действующей обратной связью. Вот как описывает эти особенности речевой коммуникации К. Черри: «Предположим, что в некоторый момент вы включились в цикл (речевого общения. — *Р. П.*) в качестве слушателя и принимаете поток звуков. В момент приема звуков вы обладаете некоторыми M_1 — „состоянием ума", которое соответствует определенным убеждениям. Можно считать, что такие убеждения составляют часть исходной системы взвешенных гипотез, как если бы мозг был физиологическим представлением функции правдоподобия в пространстве большого числа измерений — в таком пространстве синтактических и семантических аспектов сигнала, какое можно построить на основе предшествующей беседы и какое зависело бы от всего прошлого опыта. Затем можно было бы считать, что вы имеете определенное исходное состояние готовности или исходное состояние убежденности». Когда ваш собеседник, продолжает К. Черри, «начинает говорить, то произносимые им звуки являются для вас данными о содержании сообщения. Что же касается хранимого в уме содержания сообщения, то к нему вы не имеете непосредственного доступа и не обладаете „истинными знаниями" о нем. Как и в любом

опыте, полученные данные могут в лучшем случае лишь изменить ваше состояние убежденности. Эти данные изменяют вашу многомерную функцию правдоподобия до состояния, которое можно назвать апостериорным состоянием убежденности или подготовкой к ответу. . .

«При приеме звуков, которые произнес ваш друг, производится селективное действие на совокупность гипотез — одни из них усиливаются, другие ослабляются. С этого момента вы становитесь другим человеком; опыт изменил „состояние готовности" к приему других символов. Вообще говоря, любой переданный вам звук или жест содержит символы, которые определяют изменение убеждений, так продолжается циклический целенаправленный процесс. . . Можно считать, что изменение состояния убежденности от исходного априорного к апостериорному связано с прагматически информационным содержанием символов» [50, стр. 263].

Короче говоря, в ходе речевого общения одни и те же лингвистические единицы постоянно изменяют свою значимость как в семантико-прагматическом, так и в синтактическом (вероятностно-статистическом) смысле.[18] Поэтому здесь необходимо использовать особый вероятностно-статистический аппарат, который мог бы стать основой при изучении выработки суждений, влияния определенных явлений и принятии рациональных решений [ср. 56]. На этой основе должна была бы быть также создана информационная теория, отличная в своих посылках от теории связи К. Шеннона [ср. 81].

До тех пор, пока такая теория и ее математический аппарат не доведены до реалистичного и рабочего уровня, мы вынуждены пользоваться вероятностно-информационным математическим аппаратом, заимствованным у физики и технических наук. Использование такого аппарата дает нам (см. § 2) среднюю условную энтропию (информацию) для каждой буквенной позиции. Иными словами, возникает некоторая усредненная схема текста, в которой уже не отражена значимость («valeur») отдельных лингвистических единиц (речь здесь идет о вероятностной оценке этой значимости).

Особого внимания заслуживает также отношение к исследуемому тексту самого испытуемого, который выступает одновременно в роли наблюдателя и в функции вычислительного устройства, оценивающего вероятности отдельных букв в каждом участке текста. Известно, что положение прибора и наблюдателя относительно исследуемого явления имеет принципиальное значение. Даже в астрономии «эффект присутствия наблюдателя при постановке эксперимента» столь существен, что, как показывает теория относительности, «пренебречь им уже невозможно» [79, стр. 97—98]. Что же касается исследования речевой коммуникации, то здесь вопрос об отношении прибора и наблюдателя к процессу общения приобретает особую остроту.

Хотя информант и поставлен по отношению к угадываемому тексту в положение пассивного наблюдателя, никак не влияющего на «состояние убежденности» автора текста, состояние убежденности и готовности к приему символов самого угадчика постоянно меняется по мере развертывания эксперимента. В начале текста информант угадывает буквы, опираясь на свои знания о сочетаемости букв, слогов, морфем и слов. Изменения состояния убежденности угадчика диктуются его лингвистическим опытом. Используя различные приемы усреднения и оценки достоверности, мы пытались соотнести индивидуальный опыт угадчика с коллективным лингвистическим опытом народа — носителя языка (ср. § 15—17).

Однако, отгадав одну-две фразы, а иногда всего лишь несколько слов, испытуемый начинает все чаще обращаться к своей собственной оценке тематики текста, а также к своему жизненному и культурному опыту.[19] Иными словами, наш эксперимент по угадыванию оценивает не только статистическую информацию текста, но также и прагматическую информацию, связанную с индивидуальным восприятием этого текста угадчиком.

[18] В связи с этим понятно, почему так трудно собирать обширную лингвистическую (равно как и социологическую, психологическую, экономическую) информацию, которая состояла бы из серий однородных данных, достаточных для строгого применения того вероятностно-статистического аппарата, который используется в естественных науках. Отсюда проистекают также трудности при определении степени достоверности количественных величин информации, характеризующих отдельные буквенные позиции и другие участки текста.

[19] В этом смысле ход угадывания букв неизвестного текста можно сравнить с процессом чтения печатного текста «по складам».

Введенная в § 19 мера контекстной обусловленности одновременно оценивает и статистическую информацию, связанную с сочетаемостью лингвистических единиц, и прагматическую информацию, характеризующую опыт угадчика. Если в начале текста доля последней настолько мала, что ею можно пренебречь, то по мере развертывания текста ее удельный вес становится все более значительным. В связи с этим возникает вопрос о том, как отделить в ходе нашего эксперимента статистическую информацию текста от прагматического информационного «шума».

Мы пытались обойти эти трудности, угадывая лишь начальные участки текста, где величина прагматической информации, как уже указывалось, еще сравнительно невелика. Мы стремились уменьшить величину прагматического шума путем разных приемов оптимизации угадывания и приближения его к идеальному предсказанию (ср. §§ 15—16), а также путем коллективного угадывания (ср. § 17).

«Чистую» оценку статистической информации, содержащейся в тексте, можно было бы получить, сопоставляя общее количество информации, извлекаемое путем эксперимента, с оценкой прагматической информации. Но для этого нужно научиться оценивать эту последнюю. Здесь, например, может оказаться полезным сопоставление величин энтропии (информации) или избыточности, получаемых при угадывании научно-делового (или вообще специального) текста специалистом в данной области знаний, с аналогичными величинами, получаемыми из эксперимента по угадыванию этого же текста неспециалистом.[20]

Информационная неоднозначность результатов нашего эксперимента относительно разных участков исследуемого текста, вынужденное применение такой методики расчета, которая плохо учитывает изменения лингвистической и вероятностно-статистической значимости лингвистических единиц, отсутствие надежных приемов для определения того, насколько достоверны результаты эксперимента

[20] Для решения этой задачи может быть использован и предложенный недавно П. Б. Непельским прием по оценке субъективной информации. Этот прием строится на сравнении энтропии распределения букв для каждого шага угадываемых текстов с энтропией распределения тех предложений, которые предлагает информант. См. об этом в материалах X Международного лингвистического съезда в Бухаресте (1967 г.), ср. также [40а].

по угадыванию, приводит к тому, что получаемые величины энтропии (информации) и избыточности следует рассматривать как грубые оценки информационных характеристик языка и текста.

Однако при всех недостатках и неясностях в технологии самого эксперимента и в используемых в настоящее время методах обработки его результатов эксперимент по угадыванию вскрывает ряд аспектов функционирования языка и речи.

Среди них наиболее важными являются следующие особенности.

1. *Лингвистический текст дает квантовое распределение информации.* Это объясняется, очевидно, тем, что письменная и устная речь воспринимается нашими органами чувств и перерабатывается соответствующими центрами мозга не непрерывно, но путем периодической отдачи накопленных порций информации (литературу по этому вопросу см. в статье [25, стр. 125, примеч.; 27]).

2. В качестве кванта статистической информации в тексте обычно выступает слово. В условиях естественного побуквенного и пословного развертывания письменного текста основная часть информации группируется в начале слова. Концы слов, а у длинных слов и средние участки часто оказываются избыточными. Такое распределение статистической информации, отражающее разветвленность продолжений слова при данном словаре и морфологии, объясняет нам не только механизм аббревиации. Оно проясняет вопрос об установлении границ слова в тексте. По всей вероятности, принимая решение о границах слова, *интерпретант* (слушающий или говорящий) опирается не на непосредственно примыкающие к этой границе комбинации букв, фонем или звуков. Он гипотетически определяет положение этих границ несколько раньше, исходя при этом из своих убеждений о возможностях тех или иных продолжений для данного начала слова. Приняв решение, что данное начало слова имеет одно возможное или весьма вероятное продолжение, интерпретант тем самым определяет, где кончается это слово.

Такой подход к определению и различению пограничных сигналов может оказаться плодотворным при решении технических вопросов автоматической сегментации

устного текста, а также машинного выделения и распознавания его единиц.

3. Тот факт, что статистическая информация неравномерно распределена по слову, имеет принципиальное значение при решении таких инженерно-лингвистических задач, как представление слов и словоформ в памяти анализирующей или синтезирующей текст электронно-вычислительной машины (ЭВМ). Поскольку середины длинных слов являются обычно избыточными для передачи содержащейся в таком слове семантической информации, достаточно фиксировать его начало и конец. При обработке такого слова на ЭВМ оно может быть введено в машину в аббревиированном виде, т. е. с указанием лишь начальных и конечных букв. В этом случае можно значительно сократить количество ячеек, необходимых для записи в памяти ЭВМ слов текста и машинного словаря [см. 14].

4. Неравномерностью распределения статистической информации в слове можно объяснить некоторые особенности исторического развития языков. Сюда относятся: фонетическое выветривание середин длинных слов, ослабление и отпадение конечных участков слов, приводящее обычно к разрушению флексий, значительная сопротивляемость фонетическим изменениям начала слова, а также препозитивных служебных форм. В этой связи проясняются процессы, приводящие к замене флективной техники аналитическим строем языка [ср. 26, 29].

5. Эксперимент по угадыванию позволяет оценить функциональную отдачу контекста и флективной морфологии (ср. § 26). Этот эксперимент описывает с помощью информационно-вероятностных измерений действие структурных связей и их иерархию в тексте. Стратегия угадчика показывает, как на комбинаторику фигур — букв и слогов — накладываются ограничения, связанные с сочетаемостью простых знаков — морфем. В свою очередь комбинаторика морфем ограничивается сочетаемостью знаков более высокого порядка — слов. Затем по мере развертывания текста на комбинаторику слов напластовываются ограничения в сочетаемости словосочетаний и предложений, а на эти последние в свою очередь накладываются экстралингвистические композиционно-сюжетные ограничения всего повествования.

ПРИЛОЖЕНИЕ

Таблица значений $k\,(q_k - q_{k+1})\log_2 k$

k	$q_k - q_{k+1}$									
	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2	0.0200	0.400	0.0600	0.0800	0.1000	0.1200	0.1400	0.1600	0.1800
3	0.0475	0.0950	0.1425	0.1900	0.2375	0.2850	0.3325	0.3800	0.4275
4	0.0800	0.1600	0.2400	0.3200	0.4000	0.4800	0.5600	0.6400	0.7200
5	0.1161	0.2322	0.3483	0.4644	0.5805	0.6966	0.8127	0.9288	1.0449
6	0.1551	0.3102	0.4651	0.6204	0.7755	0.9306	1.0857	1.2408	1.3959
7	0.1965	0.3930	0.5895	0.7860	0.9825	1.1790	1.3755	1.5720	1.7685
8	0.2400	0.4800	0.7200	0.9600	1.2000	1.4400	1.6800	1.9200	2.1600
9	0.2853	0.5706	0.8559	1.1412	1.4265	1.7118	1.9971	2.2824	2.5677
10	0.3322	0.6644	0.9966	1.3288	1.6610	1.9932	2.3254	2.6576	2.9898
11	0.3805	0.7610	1.1415	1.5220	1.9025				
12	0.4302	0.8604	1.2906	1.7208	2.1510				
13	0.4811	0.9622	1.4433	1.9244	2.4055				
14	0.5330	1.0660	1.5990	2.1320	2.6650				
15	0.5860	1.1720	1.7580	2.3440	2.9300				
16	0.6400	1.2800	1.9200	2.5600	3.2000				
17	0.6949	1.3898	2.0847	2.7796	3.4745				
18	0.7506	1.5012	2.2518	3.0024	3.7530				
19	0.8071	1.6142	2.4213	3.2284	4.0355				
20	0.8644	1.7288	2.5932	3.4576	4.3220				
21	0.9224	1.8448	2.7672						

Продолжение

k	$q_k - q_{k+1}$								
	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
22	0.9811	1.9622	2.9433	3.0024	3.7530				
23	1.0404	2.0808	3.1212	3.2284	4.0355				
24	1.1004	2.2008	3.3012	3.4576	4.3220				
25	1.1610	2.3220	3.4830						
26	1.2222	2.4444	3.6666						
27	1.2838	2.5676	3.8514						
28	1.3460	2.6920	4.0380						
29	1.4088	2.8176	4.2264						
30	1.4721	2.9442	4.4163						
31	1.5358	3.0716	4.6074						
32	1.6000	3.2000	4.8000						
33	1.6646	3.3292	4.9938						

СОКРАЩЕНИЯ

ВЯ — Вопросы языкознания. Москва.
дв. ед. — двоичные единицы, в двоичных единицах.
ИП — сб. «Инженерная психология». Москва, 1964.
ЛГУ — Ленинградский государственный университет.
МГПИИЯ — Минский государственный педагогический институт иностранных языков.
НЛ — сб. «Новое в лингвистике». Москва.
ПК — сб. «Проблемы кибернетики». Москва.
РТИК — Кг. Шеннон. Работы по теории информации и кибернетике. Москва, 1963.
СтР — сб. «Статистика речи». Москва—Ленинград, 1968.
ТВП — Теория вероятностей и ее применения. Москва.
УМН — Успехи математических наук. Москва.
ЭВМ — электронно-вычислительная машина.
ЭЯСР — сб. «Энтропия языка и статистика речи». Минск, 1966.

BSTJ — Bell System Technical Journal.
SCL — Studii şi cercetări lingvistice. Bucureşti.

ЛИТЕРАТУРА

1. П. М. Алексеев. Частотный словарь английского подъязыка электроники. Автореф. канд. дисс. Л., 1965.
2. Х. З. Багирова, Д. А. Байтанаева, К. Б. Бектаев. Энтропия и избыточность двух тюркских языков (предварительные данные). ЭЯСР.
3. К. Б. Бектаев, Л. Е. Машкина, Т. А. Микерина, А. С. Ротарь. Энтропия словоформы и словосочетания в английских и немецких текстах. ЭЯСР.
4. Г. Г. Белоногов. О некоторых статистических закономерностях в русской письменной речи. ВЯ, 1962, № 1.
5. Г. П. Богуславская. Информационно-статистическое строения английской словоформы. ЭЯСР.
6. Г. П. Богуславская. О воздействии лексико-грамматического контекста на информационное строение английской словоформы. МГПИИЯ. Материалы XVIII научно-теоретической конференции (20—25 февраля 1966 г.). Тезисы докладов, 1966.
7. Г. П. Богуславская. О теоретико-информационном исследовании английского языка и его жанрово-стилистических разновидностей (энтропия английской письменной речи). МГПИИЯ. XVII научная сессия, посвященная итогам научно-исследовательской работы за 1964 г., II, 1964.
8. Г. П. Богуславская. Энтропия английского печатного текста. Автореф. канд. дисс. Минск, 1966.
9. М. М. Бонгард. О понятии «полезная информация». ПК, вып. 9, 1963.
10. М. Л. Гаспаров. Статистическое обследование русского трехударного дольника. ТВП, т. VIII, 1963, вып. 1.
11. А. В. Гут. Энтропия польского печатного текста. Дипломная работа, филфак ЛГУ. Л., 1966.
12. Р. Л. Добрушин. Упрощенный метод экспериментальной оценки энтропии статистической последовательности. ТВП, т. III, 1958, вып. 4.
13. Л. И. Ешан. Опыт статистического описания научно-технического стиля румынского языка (на материале текстов по радиоэлектронике). Автореф. канд. дисс. Л., 1966.
14. А. В. Зубов, К. Ф. Лукьянцепков, Р. Г. Пиотровский, Э. Н. Хотяшов. Статистическое описание текста с помощью ЭВМ. СтР.
15. Р. А. Казарян. Оценка энтропии армянского текста. Известия Академии наук Армянской ССР, Физико-математические науки, XIV, 1961, вып. 4.

16. А. Н. Колмогоров. Три подхода к определению понятия «количество информации». Сб. «Проблемы передачи информации», т. I, вып. 1, М., 1965.
17. И. А. Короленко. И. В. Матковский, Л. А. Новак, Р. Г. Пиотровский. Информационная структура слова. Сб. «Методы сравнительно-сопоставительного изучения современных романских языков», М., 1966.
18. И. А. Короленко, И. В. Матковский, Л. А. Новак, Р. Г. Пиотровский. Кодовые оценки типологии романских языков (энтропия румынских и молдавских текстов). сб. «Координационное совещание по сравнительному и типологическому изучению романских языков», Л., 1964.
19. Д. Н. Ленской. Об оценке энтропии адыгейских печатных текстов. (Предварительные опыты). Ученые записки Кабардино-балкарского университета, вып. XVI. Серия физико-математическая, Нальчик, 1962.
20. В. В. Макаров, Л. А. Новак, Р. Г. Пиотровский. Статистико-информационные измерения родства языков. Тезисы III Всесоюзного совещания романистов, Минск, 1967.
21. Л. А. Новак. Частотный словарь молдавского и румынского языков. Автореф. канд. дисс. Л., 1962.
22. Е. В. Падучева. Статистическое исследование структуры слога. Сб. «Вопросы статистики речи». Л., 1958.
23. Н. В. Петрова. Кодовые характеристики письменного текста. СтР.
24. Н. В. Петрова, Р. Г. Пиотровский. Слово, контекст, морфология. ВЯ, 1966, № 2.
25. А. А. Пиотровская, Р. Г. Пиотровский, К. А. Разживин. Энтропия русского языка. ВЯ, 1962, № 6.
26. Р. Г. Пиотровский. Аналитизм и его вероятностно-информационные механизмы. Сб. «Аналитические конструкции в языках различных типов», М.—Л., 1965.
27. Р. Г. Пиотровский. Информационные измерения печатного текста. ЭЯСР.
28. Р. Г. Пиотровский. Коллектив «статистика речи». Сб. «Межвузовская конференция по вопросам частотных словарей и автоматизации лингвистических работ». Тезисы докладов и сообщений, Ленинград, 18—20 октября 1966 г., Л., 1966.
29. Р. Г. Пиотровский. О теоретико-информационных параметрах устной и письменной форм языка. Сб. «Проблемы структурной лингвистики», М., 1962.
30. Р. Г. Пиотровский. Очерки по стилистике французского языка. 2-е изд., Л., 1960.
31. Р. Г. Пиотровский. Теоретико-информационная структура русского слова. Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung, Bd. 18, Hft. 2, 1965.
32. Р. Г. Пиотровский, П. М. Алексеев, Е. А. Чернядьева. Статистика речи и закономерности языка. Тезисы докладов межвузовской конференции на тему

«язык и речь» (27 ноября—1 декабря 1962 г. 1-й Московский гос. педагогический институт иностр. языков), 1962.
33. А. П. Савчук. Об оценках энтропии языка по Шеннону. ТВП, т. IX, 1964, вып. 1.
34. А. П. Савчук. Экспериментальное определение энтропии русского языка. Научное совещание, посвященное применению математических методов в изучении языка художественных произведений, Горький, 1961.
35. В. Ю. Урбах. К учету корреляции между буквами алфавита при вычислении количества информации в сообщении. ПК, вып. 10, 1963.
36. В. К. Фадеев. К понятию энтропии конечной вероятностной схемы. УМН, XI, 1956, вып. 1.
37. А. А. Харкевич. О ценности информации. ПК, вып. 4, 1960.
38. А. Я. Хинчин. Понятие энтропии в теории вероятностей. УМН, VIII, вып. 3, 1953.
39. А. М. Яглом и И. М. Яглом. Вероятность и информация. 2-е изд., М., 1960.
40. M. Aborn, H. Rubinstein and T. D. Sterling. Sources of Contextual Constraints upon Words in Sentences. Journal of Experimental Psychology, LVII, 3, 1959.
40a. F. Attneave. Applications of Information Theory to Psychology. New York, 1959.
41. Y. Bar-Hillel. An Examination of Information Theory. Philosophy of Science, XXXII, 2, 1955.
42. Y. Bar-Hillel. Logical Syntax and Semantics. «Language», XXX, 2, 1954.
43. Y. Bar-Hillel. Semantic Information and its Measures. Cybernetics-Circular Causal and Feedback Mechanisms in Biological and Social Systems. «Transactions of the Tenth Conference», New York, 1955.
44. G. A. Barnard. Statistical Calculation of Word Entropies for Four Western Languages. «Transactions of the Institute of Radio Engineers on the Information Theory», I, 1, 1955.
45. L. Brillouin. Science and Information Theory. New York, 1956 (цит. по русск. переводу: Л. Бриллюэн. Наука и теория информации. М., 1960).
46. N. G. Burton, J. C. Licklider. Long-range Constraints in Statistical Structure of Printed English. American Journal of Psychology, LXVIII, 4, 1955.
47. R. Carnap. Logical Foundations of Probability. Chicago, 1950.
48. R. Carnap and Y. Bar-Hillel. An Outline of a Theory of Semantic Information. Massachusetts Institute of Technology, Research Laboratory of Electronics, Technical Report, № 247, 1953, October 27.
49. D. H. Carson. Letter Constraints within Words in Printed Englich. «Kybernetik», I, 1, 1961.
50. C. Cherry. On Human Communication, London, 1961 (цит. русск. перевод одной из глав этой книги: К. Черри. О логике связи (синтактика, семантика и прагматика). ИП).
51. L. Doležel. Předběžný odhad entropie a redundance psané češtiny. «Slovo a Slovesnost», XXIV, 3, 1963.

52. P. Elias. Predictive Coding. «Transactions of the Institute of Radio Engineers», vol. I, March, 1955 (цит. по русск. переводу: П. Элиас. Кодирование с предсказанием. Сб. «Теория информации и ее приложения», М., 1959).
53. A. Feinstein. Foundations of Information Theory. New York—Toronto—London, 1958 (цит. по русск. переводу: А. Файнстейн. Основы теории информации. М., 1960).
54. B. L. Fritz, G. W. Grier. Pragmatic Communication. «Information Theory in Psychology», Illinois, 1955.
55. F. C. Fritz, W. H. Sumby. Control tower Language. Journal of the Acoustical Society of America, XXIV, 6, 1952.
56. I. J. Good. Probability and the Weighing of Evidence. London, 1950.
57. G. Herdan. Quantitative linguistics. London, 1964.
58. L. Hjelmslev. Prolegomena to a Theory of Language. Baltimore, 1953 (цит. по русск. переводу: Л. Ельмслев. Пролегомены к теории языка. НЛ, вып. 1, 1960).
59. Ch. F. Hockett. Grammar for the Hearer. «Structure of Language and its Mathematical Aspects» (Proceedings of Symposia in Applied Mathematics, vol. XII), 1961 (цит. по русск. переводу: Ч. Хоккет. Грамматика для слушающего. НЛ, вып. IV, 1965).
60. A. M. Jaglom und I. M. Jaglom. Warscheinlichkeit und Information. Berlin, 1965 (дополненный немецк. перевод работы [39]).
61. H. Karlgren. Information Estimates. Proceedings of the IXth International Congrees of Linguists, London—The Hague—Paris, 1964.
62. A. Korzybski. Science and Sanity. Lancaster (Rens), 1941.
63. K. Kupfmüller. Die Entropie der deutschen Sprache. Fernmeldetechnische Zeitschrift, VII, 6, 1964.
64. J. Marschak. Towards an Economic Theory of Information and Organisation. «Decision Processes», New York, 1954.
65. G. A. Miller. The Magical Number Seven. Plus or Minus Two: Some Limits on Our Capacity for Processing Information. Psychological Review, 63, № 2, 1956 (цит. по русск. переводу: Дж. А. Миллер. Магическое число семь плюс или минус два. О некоторых пределах нашей способности перерабатывать информацию. ИП).
66. Ch. W. Morris. Foundations of the Theory of Signs. International Enciclopedia of Unified Science, vol. I, № 2, University of Chicago Press, Chicago, 1938.
67. E. B. Newman and L. Gerstman. A New Method for Analysing Printed English. Journal of the Experimental Psychology, XLIV, 1952.
68. E. W. Newman and N. C. Waugh. The Redundancy of Texts in Three Languages. «Information and Control», III, 2, 1960.
68a. L. Novac, R. Piotrovski. Experimentul de predicție și entropia limbii române. SCL, 1968 (в печати).
69. N. Petrova, R. Piotrovski, R. Giraud. L'entropie du français écrit. Bulletin de la société de linguistique de Paris, LIX, 1964.

70. R. G. Piotrovsky. The Place of Information-Carrying Elements in a Word. «Abstracts of the Conference on Mathematical Liguistics», April 15—21, 1959. Soviet Developments, Wachington, August 31, 1959. JPRS: 893—D.
71. W. van O. Quine. From a Logical Point of View. Harvard Univ. Press, Cambridge (Mass.), 1953.
72. A. Rapoport. What is Semantics? «Language, Meaning and Maturity». Selections from «ETC»: A Review of General Semantics 1943—1953, ed. by S. I. Haykawa, New York, 1954.
73. K. E. Riegel and R. M. Riegel. Prediction of Word-Recognition threshold on the Basis of Stimulus-Parameters. «Language and Speech», IV, 3, 1961.
74. I. Rhodes. The National Bureau of Standards Method of Syntactic Integration. «Proceedings of the National Symposium of the Machine Translation», Los Angelos, 1960, New York, 1961.
75. A. Schaff. Wstęp do semantyki. Warszawa, 1960 (цит. по русск. переводу: А. Шаф ф. Введение в семантику. М., 1963.)
76. C. E. Shannon. A Mathematical Theory of Communication. BSTJ, XXVII, 3, 1948 (цит. по русск. переводу: К. Шеннон. Математическая теория связи. РТИК).
77. C. E. Shannon. Prediction and Entropy of Printed English. BSTJ, XXX, 1, 1951 (цит. по русск. переводу: К. Шеннон. Предсказание и энтропия печатного английского текста. РТИК).
78. R. Wells. A Mesure of Subjective Information. «Structure of Language and its Mathematical Aspects». Proceedings of Symposia in Applied Mathematics, vol. XII, 1961 (цит. по русск. переводу: Р. Уэллз. Мера субъективной информации. ИЛ, вып. IV, 1965).
79. N. Wiener. God and Golem. A Comment on Certain Point where Cybernetics impinges on Religion, The MIT Press. Massachusetts Institute of Technology, Cambridge (Mass.), 1964 (цит. по русск. переводу: Н. Винер. Творец и робот. Обсуждение некоторых проблем, в которых кибернетика сталкивается с религией. М., 1966).
80. L. Wittgenstein. Tractatus Logico-Philosophicus. London, 1933 (цит. по русск. переводу: Л. Витгенштейн. Логико-философский трактат. М., 1958).
81. D. M. Mac Kay. Quantal Aspects of Scientific Information. Philosophical Magasine. XLI, 1950.
82. VI. Matković. Primjena teorje komunikacije na odreotivanje entropije hrvatskoga jezika Zagreb, 1957.
82a. V. Drozen, S. Langer. Statistický odhad sémantické informace. Kybernetika, 1966, № 2.
83. Ю. А. Шрейдер. О семантических аспектах теории информации. Сб. «Информация и кибернетика», М., 1967.
84. П. Б. Невельский, Р. Г. Пиотровский. О применении метода угадывания лингвистического текста в психологии. «Тезисы III съезда Общества психологов», Киев, 1968.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

ОГЛАВЛЕНИЕ

Раймонд Генрихович Пиотровский

ИНФОРМАЦИОННЫЕ ИЗМЕРЕНИЯ ЯЗЫКА

Утверждено к печати Институтом языкознания АН СССР

Редактор Издательства *А. А. Зырин*
Технический редактор *Г. А. Бессонова*. Корректор *Л. Б. Панкова*

Сдано в набор 4/VIII 1967 г. Подписано к печати 19/IV 1968 г. РИСО АН СССР № 22-134В. Формат бумаги 84×108 1/32. Бум. л. 1 7/8. Печ. л. 3 3/8 + 1 вкл. (1/8 п. л.) = 6,3 усл. печ. л. Уч.-изд. л. 6,7. Изд. № 3187. Тип. зак. № 460. М-11508. Тираж 3500. Бумага типографская № 1. *Цена 42 коп.*

Ленинградское отделение издательства «Наука»
Ленинград, В-164, Менделеевская лин., д. 1

1-я тип. издательства «Наука». Ленинград, В-34, 9 линия, д. 12

ИСПРАВЛЕНИЯ

Страница	*Строка*	*Напечатано*	*Должно быть*
59	Табл. 9, заголовок	энтропии	информации
65	7—6 снизу	коллективных	контекстных
105	Табл. (Приложение), графа 3 слева, строка 1 сверху	0,400	0,0400

Р. Г. Пиотровский