при этом детерминированно; равенство $H(P)=\log_2 m$ достигается при равновероятном появлении $a_i \in A$ — ситуация наибольшей неопределённости. При $m=2$ и равномерном появлении букв a_1 и a_2 энтропия максимальна и $H(P)=1$. Эта величина — неопределённость при равновероятном выборе из двух альтернатив используется как единица кол-ва энтропии — 1 бит.

Пусть, далее, канал работает в r-буквенном алфавите и $r<m$. Кодирование при этом будет заключаться в сопоставлении каждому символу $a_i \in A_m$ слова $b(a_i)$ в алфавите B_r.

Каждый способ кодирования характеризуется ср. числом $L(P)$ букв выходного алфавита, приходящихся на одну букву входного алфавита A_m. Для алфавитного кодирования $L(P)=\sum_{i=1}^{m} p_i l_i$, где l_i — длина слова $b(a_i)$ в алфавите B_r. Если кодирование взаимно однозначно, то

$$L(P) \geqslant H_r(P) = \sum_{i=1}^{m} p_i \log_r (1/p_i).$$

Величина $I(P)=L(P)-H_r(P)$ наз. избыточностью кодирования при распределении P. Задача состоит в отыскании в заданном классе взаимно однозначных кодирований кодирования, обладающего мин. величиной $I(P)$. Существование минимума и его значение устанавливаются теоремой Шеннона для канала без шума, гласящей, что для источника с конечным алфавитом A_m с энтропией $H(P)$ можно так приписать кодовые слова буквам источника, что ср. длина кодового слова $L(P)$ будет удовлетворять условиям

$$\frac{H(P)}{\log_r m} \leqslant L(P) < \frac{H(P)}{\log_r m} + 1.$$

Оптимальным считается такой код, что никакой другой не обеспечит меньшего значения $L(P)$.

Конструктивная процедура отыскания оптим. кода для кодирования данного множества сообщений предложена в 1952 Д. Хафменом (D. R. Huffman). Идея заключается в том, что буквы алфавита A_m упорядочиваются по вероятности и более вероятным приписываются более короткие кодовые слова. Код Хафмена обладает след. свойствами: слово, соответствующее наименее вероятному сообщению, имеет наибольшую длину; два наименее вероятных сообщения кодируются словами одинаковой длины, одно из к-рых оканчивается нулём, а другое — единицей ($r=2$).

Оптимальное равномерное кодирование. Пусть источник с двухбуквенным алфавитом $A_2=\{0, 1\}$ и $P=\{q, p\}$ генерирует слова длиной l. Относительно всего множества из 2^l слов (словаря источника) существует утверждение, что при $p \neq q$ и достаточно больших l словарь источника распадается на два подмножества: группу из $2^{lH(P)}$ равновероятных слов (рабочий словарь источника) и группу слов с суммарной вероятностью, близкой к нулю («нетипичные» последовательности). Здесь $H(P)$ — энтропия на символ источника. Доля слов рабочего словаря весьма мала и с увеличением l стремится к нулю. Идея равномерного, или блокового, кодирования заключается в том, что кодер, получая на входе слова источника, сопоставляет кодовые слова лишь словам из рабочего словаря, кодируя все остальные одним словом, имеющим смысл ошибки. Вероятность ошибки может быть произвольно уменьшена увеличением длины слова источника. При этом объём кодируемых слов $2^{lH(P)}$ требует $n \geqslant lH(P)$ символов кодового слова. Поскольку слова рабочего словаря практически равновероятны, равновероятны будут и кодовые слова, а энтропия на символ кодового слова будет близка к 1 биту. Кодер, т. о., выдаёт слова длиной $n<l$, экономя за счёт того, что «догружает» каждый символ до максимально возможной информационной нагрузки в 1 бит.

Кодирование источника приобретает новое значение в связи с необходимостью «сжатия» информационных массивов данных в базах и банках данных. Массивы организационной, экономич., измерит. информации имеют столь большую избыточность, что допускают сжатие, доходящее до 80—85%. Развитые системы управления базами данных (СУБД) имеют спец. программы (утилиты) анализа, сжатия и восстановления текста, работающие на принципах, изложенных выше.

Корректирующее кодирование информации. Его целью является обнаружение и (или) исправление ошибок в кодовых словах, возникших при передаче информации по каналу с шумом. Коррекция искажений возможна за счёт введения избыточности в систему передачи. При этом из всего множества слов кодера канала N_0 лишь N будет соответствовать передаваемым сообщениям (разрешённые слова). Теоретически при этом доля обнаруженных ошибок не превысит $1-N/N_0$.

Предполагается, что информационное слово $U=(u_1, \ldots, u_n)$, где $u_j=0, 1$, поступает на вход кодера канала (в дальнейшем — кодера), ставящего ему в соответствие кодовое слово $X(x_1, \ldots, x_l)$, $x_i=\overline{0,1}$, $l>n$. Кодер, т. о., добавляет по определ. правилу к слову U группу из $k=l-n$ избыточных (корректирующих) разрядов. Кодовое слово X поступает в канал с шумом, где помеха искажает нек-рые из символов x_i. Принятое на выходе канала слово $Y=(y_1, \ldots, y_l)$ поступает на декодер, восстанавливающий (с нек-рым приближением) слово X. С кодовыми словами оперируют как с векторами в линейном векторном пространстве с метрикой Хэмминга, задающей расстояние между векторами X' и X''

$$d(X', X'') = \sum_{i=1}^{l} (x_i' \oplus x_i'').$$

Теорема Шеннона для каналов с шумом, утверждающая, что при помощи подходящих кодов можно передавать информацию так, чтобы вероятность ошибки после декодирования была произвольно малой при условии, что скорость передачи не превосходит пропускной способности канала связи, неконструктивна: она не указывает способа построения кода. При конструировании кода решающее значение имеет выбор модели возникновения ошибок в передаваемом слове.

Наиб. распространена модель симметричного канала с равновероятными ошибками разл. типов — перехода, напр., символа 0 в 1 и 1 в 0.

Специфична модель канала «со стиранием». Выходной алфавит такого канала содержит спец. символ стирания, в к-рый и переходят символы входного алфавита при возникновении ошибки подобного типа.

Выдвигаются разл. предположения относительно распределения ошибок в передаваемой последовательности символов (кодовом слове). Возможна модель независимых ошибок (канала без памяти), модель сгруппированных ошибок (пачек ошибок), ошибок, расположенных на определ. расстоянии друг от друга, и т. д. Распространены предположения о предельной кратности ошибок в кодовых словах [3].

В рамках последнего предположения корректирующая способность кода оценивается числом ошибок, обнаруживаемых и (или) исправляемых с его помощью в кодовых словах. Предполагается, что в канале с X посимвольно суммируется (по mod 2) шумовой вектор Z, образуя слово $Y=X \oplus Y$. Кратность возникающей в результате ошибки совпадает с числом единиц (весом Хэмминга) в Z. В векторе из l элементов не более чем r единиц могут быть размещены $\sum_{m=1}^{r} C_l^m$ способами. Это — то разнообразие ошибок, к-рое может возникнуть при передаче.

Основной характеристикой кода, определяющей его корректирующую способность по отношению к независимым ошибкам, является кодовое расстояние. Кодовое расстояние является наименьшим хэмминговым