Кодирование источника (эффективное кодирование)

Реальные источники редко обладают максимальной энтропией, поэтому их принято характеризовать так называемой избыточностью

$$\kappa = \frac{H_{\max} - H}{H_{\max}}$$

Для независимых источников (источников без памяти) избыточность равна нулю (а энтропия максимальна) при равновероятности символов.

Кодирование источника

Для источников с памятью *избыточность тем больше*, чем выше *степень статистической (вероятностной) зависимости* символов в сообщении, при этом *неопределенность* относительно очередного символа в сообщении уменьшается, соответственно уменьшается и количество информации, переносимое этим символом.

$$q\square \rightarrow qu$$

Кодирование

Один символ источника → кодовое слово (кодовая комбинация)

Если для всех символов источника длина кодовых слов одинакова, код называют *равномерным*, в противном случае – *неравномерным*.

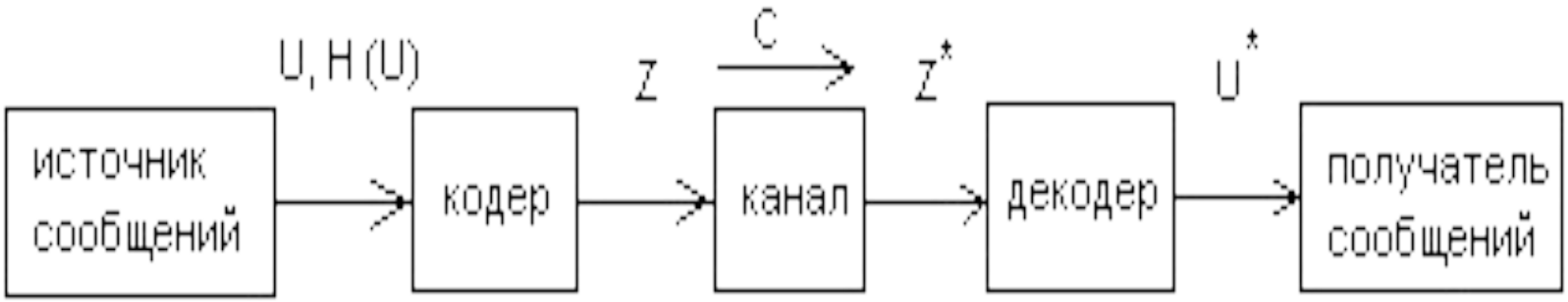
Jean-Maurice-Émile Baudot, 1845 —1903

Примером равномерного кода является код *Бодó*:

каждая из букв алфавита представляется двоичным числом фиксированной разрядности (например, а → 00001, б → 00010 и т.д.). Тогда для алфавита из 32 символов достаточно 5-значного кода.

При *не*равномерном коде говорят о *средней* длине кодового слова (усреднение длин кодовых слов производится по априорному распределению вероятностей символов источника):

$$\mu = \sum_{i=1}^{n} p_i \mu_i$$

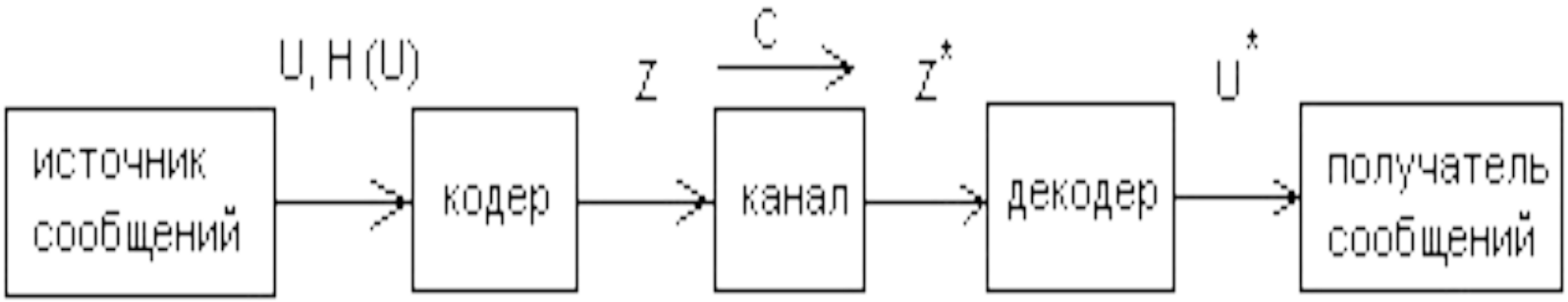


$$H'(U) = u_C \; H(U) = u_K \log M = C$$

$$\mu = u_K / u_C = H(U) / \log M$$

где H(U) - энтропия источника передаваемых сообщений, u_K и u_C - средние количества символов соответственно сообщения и кода передаваемых в единицу времени

$\mu = u_K / u_C$ - среднее количество символов кода приходящиеся на одно сообщение



$\mu = u_K / u_C$ - среднее количество символов кода приходящиеся на одно сообщение

Степень приближения к точному выполнению этих равенств зависит от степени уменьшения избыточности источника сообщений.

Кодирование позволяющее устранять избыточность источников сообщений называется ***эффективным*** *или* ***статистическим*****.**

Избыточность дискретных источников обуславливается двумя причинами:

1) памятью источника;

2) неравномерностью сообщений.

Универсальным способом уменьшения избыточности обусловленной памятью источника является укрупнение элементарных сообщений.

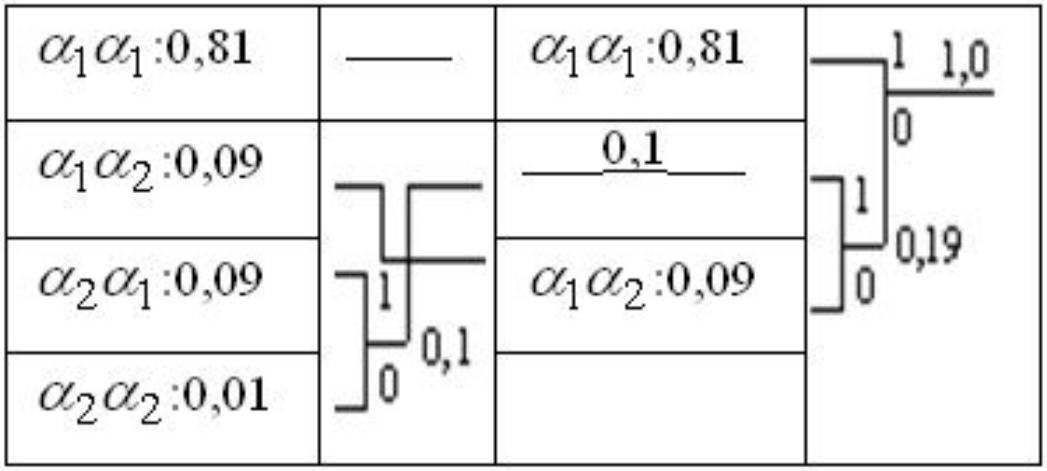
При этом кодирование осуществляется длинными блоками. Вероятностные связи между блоками меньше чем между отдельными элементами сообщений и чем длиннее блоки, тем меньше зависит между ними

Пример. Рассмотрим источник, вырабатывающий два независимых символа с вероятностями 0,1 и 0,9. В этом тривиальном случае символы алфавита кодируются символами «0» и «1». Энтропия источника H=0,469, средняя длина кодового слова равна 1,

избыточность источника и избыточность кода одинаковы и равны

$$\kappa = \frac{H_{\max}(\mathrm{A}) - H(\mathrm{A})}{H_{\max}(\mathrm{A})} = \frac{1 - 0,469}{1} = 0,531$$

Кодирование группами



$$\alpha_1\alpha_1 \to 1$$
$$\alpha_1\alpha_2 \to 00$$
$$\alpha_2\alpha_1 \to 011$$
$$\alpha_2\alpha_2 \to 010$$

$$\mu = \frac{0,81 + 0,09 \cdot 2 + 0,09 \cdot 3 + 0,01 \cdot 3}{2} = 0,645$$

$$p(1) = \frac{0,81 + 0,09 \cdot 2 + 0,01}{0,645 \cdot 2} = 0,775$$

$$H_k = -0,225 \log 0,225 - 0,775 \log 0,775 = 0,769$$

$$\kappa = \frac{H_{k\max} - H_k}{H_{k\max}} = \frac{1 - 0,769}{1} = 0,231$$

1-я Теорема Шеннона

Предельные возможности статистического кодирования раскрывается в теореме Шеннона для канала без шума, которая является одним из основных положений теории информации.

Пусть источник сообщений имеет производительность H'U) = $u_C \times H(U)$, а канал имеет пропускную способность $C = u_K \times \log M$. Тогда можно закодировать сообщения на выходе источника таким образом, чтобы получить среднее число кодовых символов приходящихся на элемент сообщения

$$\mu = u_K / u_C = (H(U)/ \log M) + \varepsilon$$

где ε- сколь угодно мало (прямая теорема)

1-я Теорема Шеннона

Получить меньшее значение μ невозможно (обратная теорема).

Обратная часть теоремы утверждающая, что невозможно получить значение
$\mu = u_K / u_C < H(U)/ \log M$

Энтропия в секунду на входе канала или производительность кодера не может превышать пропускную способность канала)

1-я Теорема Шеннона

«Основная теорема о кодировании в отсутствие шумов»:

Среднюю длину кодовых слов для передачи символов источника A *при помощи кода с основанием* m *можно как угодно приблизить к величине* H(A) / log m.

теорема определяет *нижнюю границу* средней длины кодовых слов

Пример. Если источник без памяти имеет объем алфавита 32, то при равновероятных символах его энтропия равна 5 битам. Согласно теореме, для двоичного кода (m=2) наименьшая средняя длина кодового слова составляет H(A) / log m = 5, следовательно, код *Бодо* является оптимальным кодом для равновероятного источника без памяти.

Текст на русском языке, например, имеет энтропию около 2,5 бит, поэтому путём соответствующего кодирования можно увеличить скорость передачи информации вдвое против пятиразрядного равномерного кода Бодо

Практическое значение теоремы Шеннона заключается в возможности *повышения эффективности* систем передачи информации (систем связи) путем применения *эффективного или экономного* кодирования (кодирования источника).

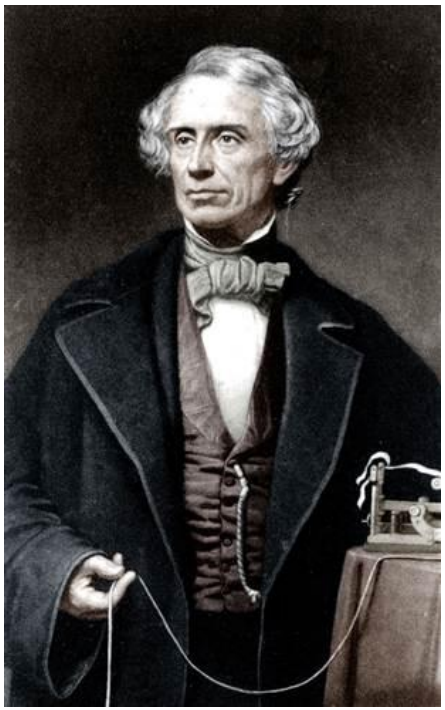
Очевидно, что экономный код должен быть в общем случае неравномерным.

Общее правило кодирования источника (без памяти) состоит в том, что более вероятным символам источника ставятся в соответствие менее длинные кодовые слова (последовательности канальных символов).

Пример 8.5. Известный код *Морзе* служит примером неравномерного кода. Кодовые слова состоят из трех различных символов: точки · (передается короткой посылкой), тире — (передается относительно длинной посылкой) и пробела (паузы).

«е» “ · ”

«ш» “ — — — — ”



Samuel Finley Breese Morse [mɔːrs] 1791 - 1872

Alfred Lewis Vail
1807 - 1859

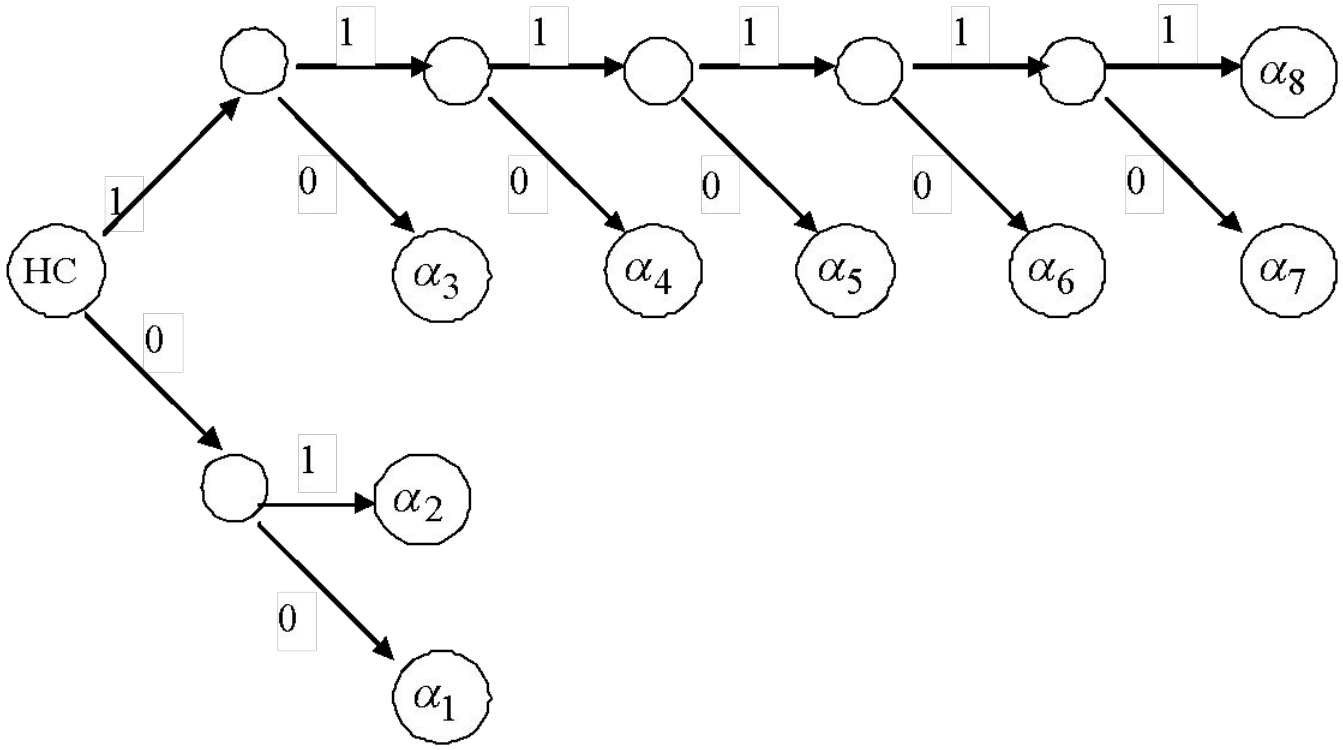
Кодирование источника по Шеннону – Фано.

Символ и его вероятность		Комбинация кодовых символов						Длина комбинации
α_i	$p(\alpha_i)$	1-й	2-й	3-й	4-й	5-й	6-й	μ_i
α_1	1/4	0	0					2
α_2	1/4	0	1					2
α_3	1/4	1	0					2
α_4	1/8	1	1	0				3
α_5	1/16	1	1	1	0			4
α_6	1/32	1	1	1	1	0		5
α_7	1/64	1	1	1	1	1	0	6
α_8	1/64	1	1	1	1	1	1	6

Ни одна кодовая комбинация не является началом какой-либо другой кодовой комбинации

(префиксное свойство)

Дерево декодирования (граф конечного автомата)



префиксные коды называют также мгновенными

Средняя длина кодовой комбинации для построенного кода

$$\mu = \sum_{i=1}^{8} p(\alpha_i)\mu_i =$$

$$= 0,75 \cdot 2 + 0,125 \cdot 3 + 0,0625 \cdot 4 + 0,03125 \cdot 5 + 0,03125 \cdot 6 = 2,469$$

Согласно теореме Шеннона при оптимальном кодировании можно достичь средней длины

$$\mu_{\min} = H(\mathrm{A}) / \log 2 = -\sum_{i=1}^{8} p(\alpha_i) \log p(\alpha_i) = 2.469$$

Заметим, что восемь различных символов источника можно представить восемью комбинациями равномерного двоичного кода (Бодо), при этом длина каждой кодовой комбинации равняется, очевидно, 3. Увеличение скорости передачи информации) составляет в данном примере около 22%.

Определим **вероятность появления определенного символа в кодовой комбинации** (пусть это будет символ 1).

Её можно найти следующим образом:

а) подсчитать количества единиц во всех кодовых словах;

б) умножить эти количества на вероятности соответствующих кодовых слов;

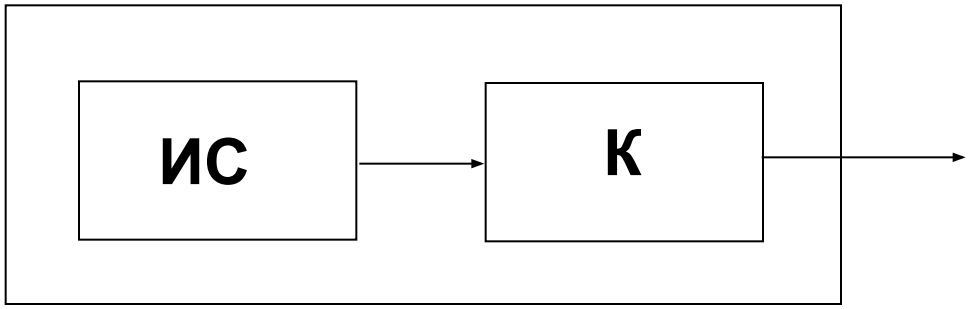
в) просуммировать полученные величины;

г) отнести результат к средней длине кодового слова.

$$p(1) = \frac{0,25 + 0,25 + 2 \cdot 0,125 + 3 \cdot 0,0625 + 4 \cdot 0,03125 + (5+6) \cdot 0,015625}{2,469} = 0,5$$

$$p(1) = \frac{0,25 + 0,25 + 2 \cdot 0,125 + 3 \cdot 0,0625 + 4 \cdot 0,03125 + (5+6) \cdot 0,015625}{2,469} = 0,5$$

при оптимальном кодировании источника кодовые символы равновероятны; такое кодирование является *безызбыточным.* Источник вместе с кодером можно рассматривать, как *новый источник с алфавитом, состоящим из кодовых символов; энтропия и избыточность этого источника – это энтропия и избыточность кода*. Оптимальный код имеет максимальную энтропию и нулевую избыточность.



Кодирование источника по Хаффману

1. Все символы алфавита записываются в порядке убывания вероятностей.

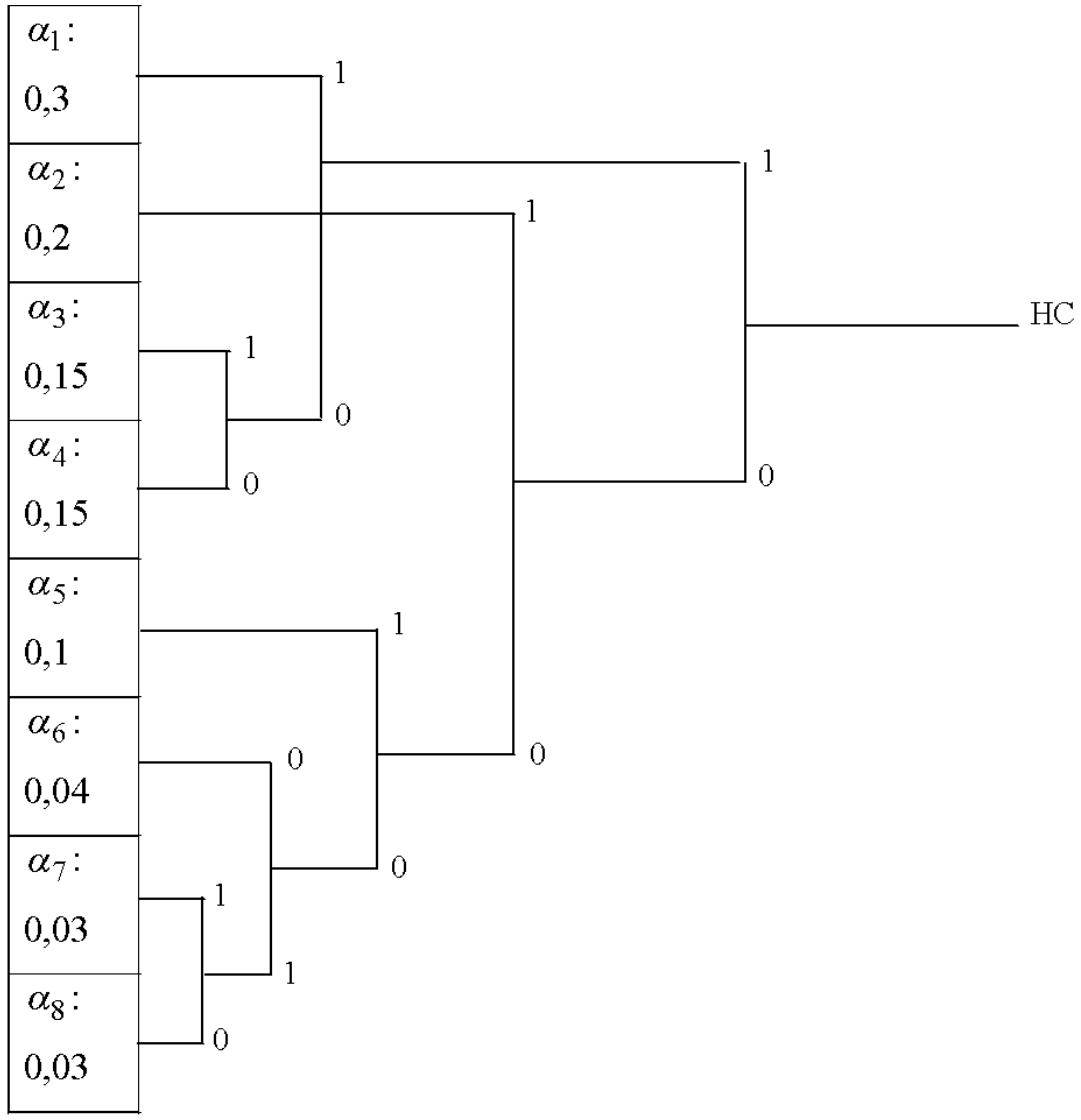
2. Два нижних символа соединяются скобкой, из них верхнему приписывается символ 0, нижнему 1 (или наоборот).

3. Вычисляется сумма вероятностей, соответствующих этим символам алфавита.

4. Все символы алфавита снова записываются в порядке убывания вероятностей, при этом только что рассмотренные символы «склеиваются», т.е. учитываются, как единый символ с суммарной вероятностью.

5. Повторяются шаги 2, 3 и 4 до тех пор, пока не останется ни одного символа алфавита, не охваченного скобкой.

Кодирование источника по Хаффмену



$$\alpha_1 \to 11$$

$$\alpha_2 \to 01$$

$$\alpha_3 \to 101$$

$$\alpha_4 \to 100$$

$$\alpha_5 \to 001$$

$$\alpha_6 \to 0000$$

$$\alpha_7 \to 00011$$

$$\alpha_8 \to 00010$$

Энтропия алфавита $H(\mathrm{A}) = 2,628$

Средняя длина кодового слова

$$\mu = \sum_{i=1}^{n} p_i \mu_i =$$

$$=0,3 \cdot 2+0,2 \cdot 2+0,15 \cdot 3+0,15 \cdot 3+0,1 \cdot 3+0,04 \cdot 4+0,03 \cdot 5+0,03 \cdot 5=2,66$$

Вероятность символа 1 в последовательности кодовых комбинаций находится как среднее количество единиц, отнесённое к средней длине кодового слова

$$p(1) = \frac{2 \cdot 0,3 + 0,2 + 2 \cdot 0,15 + 0,15 + 0,1 + 2 \cdot 0,03 + 0,03}{2,66} = 0,541$$

Избыточность кода равна 0.005

Оптимальность кода Шеннона – Фано в рассмотренном примере объясняется специально подобранными вероятностями символов, так, что на каждом шаге вероятности делятся ровно пополам.

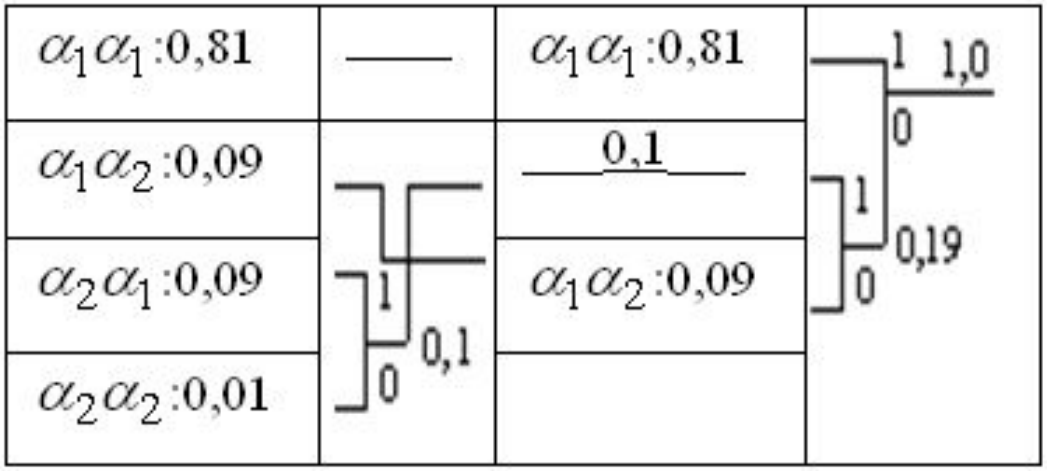
В примерах предполагались источники без памяти. В реальных сообщениях на естественных языках символы не являются независимыми; в таких случаях следует кодировать не отдельные символы (буквы), а группы букв или слова. Это уменьшает зависимость и повышает эффективность кода.

Кодирование групп символов вместо отдельных символов также повышает эффективность кодирования в случае независимого источника с сильно различающимися вероятностями символов, как это видно из следующего примера.

Пример. Рассмотрим источник, вырабатывающий два независимых символа с вероятностями 0,1 и 0,9. В этом тривиальном случае методы кодирования Хаффмена и Шеннона – Фано приводят к одинаковому коду: символы алфавита кодируются символами «0» и «1». Энтропия источника H=0,469, средняя длина кодового слова равна 1, избыточность источника и избыточность кода одинаковы и равны

$$\kappa = \frac{H_{\max}(\mathrm{A}) - H(\mathrm{A})}{H_{\max}(\mathrm{A})} = \frac{1 - 0,469}{1} = 0,531$$

Кодирование группами



$$\alpha_1\alpha_1 \to 1$$
$$\alpha_1\alpha_2 \to 00$$
$$\alpha_2\alpha_1 \to 011$$
$$\alpha_2\alpha_2 \to 010$$

$$\mu = \frac{0,81 + 0,09 \cdot 2 + 0,09 \cdot 3 + 0,01 \cdot 3}{2} = 0,645$$

$$p(1) = \frac{0,81 + 0,09 \cdot 2 + 0,01}{0,645 \cdot 2} = 0,775$$

$$H_k = -0,225 \log 0,225 - 0,775 \log 0,775 = 0,769$$

$$\kappa = \frac{H_{k\max} - H_k}{H_{k\max}} = \frac{1 - 0,769}{1} = 0,231$$