Язык и алфавит

10 класс

Как записать информацию?

Для того чтобы хранить и передавать информацию, её необходимо как-то зафиксировать, например записать с помощью символов (знаков) на каком-то языке.

Язык — это система знаков, используемая для хранения, передачи и обработки информации.

Египетское письмо	
	Рука
	Дом
	Кобра
	Лев
	Вода
	Рот
	Мужчина
	Женщина

Иероглифы (Китай)	
日	Солнце
月	Луна
雨	Дождь
山	Гора
马	Лошадь
鱼	Рыба
人	Человек
女	Женщина

- **Естественные языки** (русский, английский и др.) сформировались в результате развития человеческого общества и используются для общения людей.
- Сначала древние люди овладели **устной речью**. Поскольку человек может издавать и различать на слух не так много звуков, он стал комбинировать их, составляя слова, каждому из которых приписывался некоторый смысл.
- Затем появилась **необходимость записывать информацию**, например, для передачи потомкам. В первое время жизненный опыт пытались зафиксировать **в виде рисунков животных и предметов, затем пиктограмм** (схематических изображений), **иероглифов** (рисунок).

В современных языках используется алфавитное письмо, где каждый знак (или сочетание знаков) обозначает некоторый звук, так что с помощью небольшого набора знаков (алфавита) можно записать любые слова устной речи.

Алфавит — это набор знаков, который используется в языке.

В алфавите русского языка 33 буквы, в английском алфавите — 26.

К алфавиту языка нужно еще отнести пробел (пропуск между словами), цифры (знаки для записи чисел), знаки препинания, скобки.

Например, **алфавит**, состоящий из 33 русских букв, 10 цифр, пробела и 12 знаков препинания (точка, запятая, точка с запятой, вопросительный и восклицательный знаки, тире, двоеточие, многоточие, кавычки, круглые скобки) **имеет мощность 56** (а если различать прописные и строчные буквы, то 89).

Мощность алфавита — это количество знаков в алфавите.

Слово — это последовательность символов алфавита, которая используется как самостоятельная единица и имеет определённое значение.

Из слов составляются предложения, каждое из которых выражает определённую законченную мысль (сообщение, информацию). В языке определяются правила построения слов (грамматика), правила построения предложений (синтаксис) и правила расстановки знаков препинания (пунктуация).

- С точки зрения **теории информации**, сообщение — это любой набор знаков некоторого алфавита. Определим, сколько различных сообщений можно построить с помощью заданного количества знаков.
- Пусть, например, алфавит состоит из **четырёх знаков**: @ # $ %. С его помощью можно записать 4 разных сообщения из одного символа: @, #, $ и %. Теперь рассмотрим сообщения из двух знаков. Первый знак можно выбрать четырьмя способами, и для каждого из них есть 4 варианта выбора второго знака. **Поэтому сообщений, состоящих из двух знаков,**

@@	#@	$@	%@
@#	##	$#	%#
@$	#$	$$	%$
@%	#%	$%	%%

Рассуждая аналогично, получим, что трёхсимвольных сообщений будет $4^3 = 64$, а четырёхсимвольных — $4^4 = 256$ и т. д.

Если алфавит языка состоит из **N** символов (имеет мощность N), количество различных сообщений длиной **L** знаков вычисляется как $\mathbf{Q = N^L}$.

Кодирование и декодирование

Что такое кодирование и декодирование?

- **Код** — набор символов (знаков) для представления информации.
- Код — система условных знаков (символов) для передачи, обработки и хранения информации (сообщения).
- **Кодирование** — это представление информации в форме, удобной для её хранения, передачи и обработки (т.е. в виде кода)
- Все множество символов, используемых для кодирования, называется *алфавитом кодирования*.
- **Декодирование** — это восстановление информационного сообщения из последовательности кодов.

- В зависимости от конкретной задачи информация может кодироваться разными способами.
- Например, фраза **«Привет, Вася!»** может быть закодирована транслитом (транслитерация): **«Privet, Vasya!».**
- Такой метод раньше часто используют в электронных письмах, в тех ситуациях, когда у одного собеседника (или у обоих) на компьютере (телефоне) нет поддержки русского языка.
- Либо сообщение можно просто перевести на английский (или какой-то другой) язык, если собеседник не знает русского языка. А можно даже зашифровать.

Шифрование — это один из способов кодирования, при котором нужно скрыть смысл сообщения от посторонних

Для **кодирования числовой информации** в разных ситуациях тоже используют разные способы. Например, число 21 можно записать как XXI (в римской системе счисления) или «двадцать один» (в финансовых документах).

Способы кодирования информации

- Для кодирования одной и той же информации могут быть использованы разные способы
- Их выбор зависит от ряда обстоятельств: цели кодирования, условий, имеющихся средств.
- Выбор способа кодирования информации может быть связан с предполагаемым способом ее обработки.

Способы кодирования информации:

Графический	• с помощью рисунков или значков
Числовой	• с помощью чисел
Символьный	• с помощью символов того же алфавита, что и текст

Двоичное кодирование в компьютере

Информация, которую обрабатывает компьютер должна быть представлена двоичным кодом с помощью двух цифр: **0** и **1**.

Эти два символа принято называть
двоичными цифрами или битами.

- **Кодирование** – преобразование входной информации в форму, воспринимаемую компьютером, т.е. двоичный код.
- **Декодирование** – преобразование данных из двоичного кода в форму, понятную человеку.

Почему двоичное кодирование?

- С точки зрения технической реализации использование двоичной системы счисления для кодирования информации оказалось **намного более простым, чем применение других способов.**
- Действительно, удобно кодировать информацию в виде последовательности нулей и единиц, если представить эти значения как **два возможных устойчивых состояния** электронного элемента:

0 – отсутствие электрического сигнала

1 – наличие электрического сигнала

- **Недостаток двоичного кодирования** – длинные коды. Но в технике **легче иметь дело с большим количеством простых элементов, чем с небольшим числом сложных.**
- Способы кодирования и декодирования информации в компьютере, в первую очередь, зависит от вида информации, а именно, что должно кодироваться: **числа, текст, графические изображения или звук.**

Двоичное кодирование текстовой информации

- Традиционно для кодирования одного символа используется количество информации = 1 байту

1 символ = 1 байт = 8 бит

- Для кодирования **одного символа** требуется **один байт** информации.
- Учитывая, что каждый бит принимает значение 1 или 0, получаем, что с помощью 1 байта можно закодировать 256 различных символов.

$2^8=256$

Кодирование заключается в том, что каждому символу ставится в соответствие уникальный двоичный код

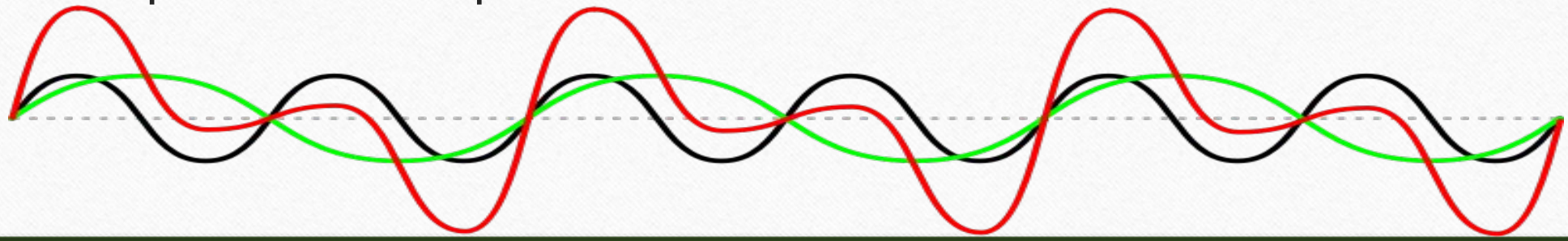
от 00000000 до 11111111

или десятичный код от 0 до 255.

Важно, что **присвоение символу конкретного кода** – это вопрос соглашения, которое фиксируется кодовой таблицей.

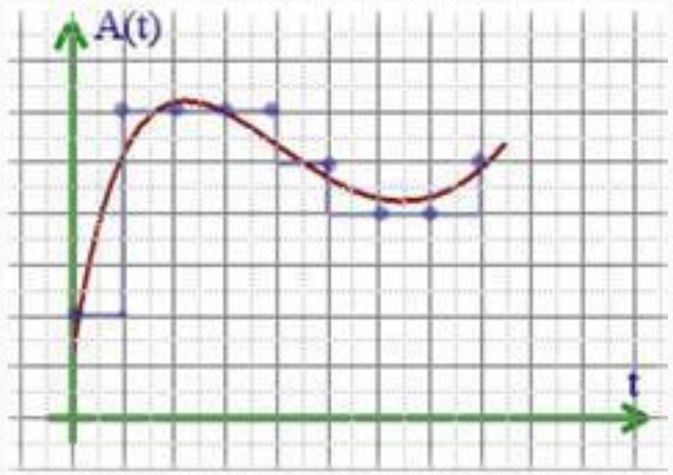
Кодирование звука

- **Звук** – волна с непрерывно изменяющейся амплитудой и частотой. Чем больше амплитуда, тем он громче для человека, чем больше частота, тем выше тон.
- Сложные непрерывные сигналы можно с достаточной точностью представлять в виде **суммы некоторого числа простейших синусоидальных колебаний.**
- Каждое слагаемое, то есть каждая синусоида, может быть точно задана некоторым набором числовых параметров – амплитуды, фазы и частоты, которые можно рассматривать как код звука в некоторый момент времени.



Временная дискретизация звука

- В процессе кодирования звукового сигнала производится его **временная дискретизация** – непрерывная волна разбивается на отдельные маленькие временные участки и для каждого такого участка устанавливается определенная величина амплитуды.
- Таким образом непрерывная зависимость амплитуды сигнала от времени заменяется на дискретную последовательность уровней громкости



- Качество двоичного кодирования звука определяется **глубиной кодирования и частотой дискретизации**.
- **Частота дискретизации** – количество измерений уровня сигнала в единицу времени.
- Количество уровней громкости определяет **глубину кодирования**.
- Современные звуковые карты обеспечивают 16-битную глубину кодирования звука.
- При этом количество уровней громкости равно $N = 2^I = 2^{16} = 65536$.

Декодирование — это восстановление информационного сообщения из последовательности кодов.

Например, закодированное сообщение

можно восстановить, используя код Морзе «в обратную сторону»: в этой строке закодирована фамилия «Петров».

- В некоторых случаях даже **при использовании неравномерного кода не требуется вводить символ-разделитель.** Для этого достаточно выполнение **условия Фано:** ни одно кодовое слово не совпадает с началом другого кодового слова.
- Такой **код** называют **префиксным**.

Код

Равномерный код

- В кодовых комбинациях содержится одинаковое число символов

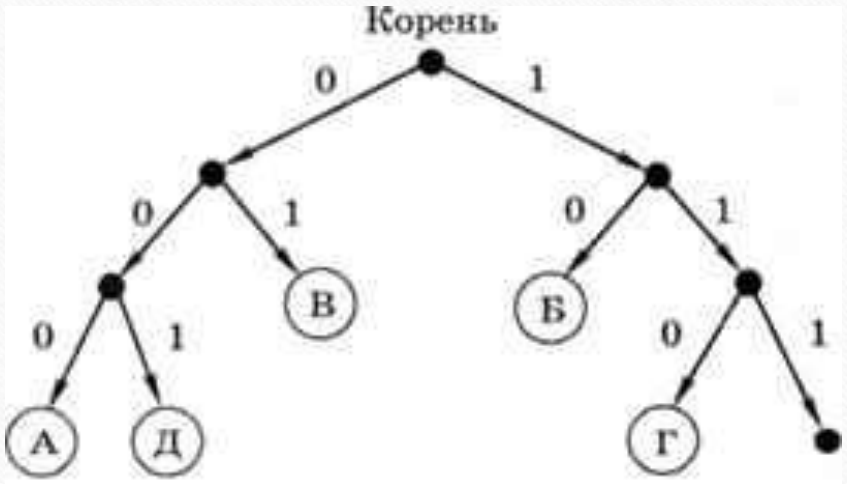
Неравномерный код

- В кодовых комбинациях содержатся разное число символов

Пример 1. Пусть для кодирования первых 5 букв русского алфавита используется таблица:

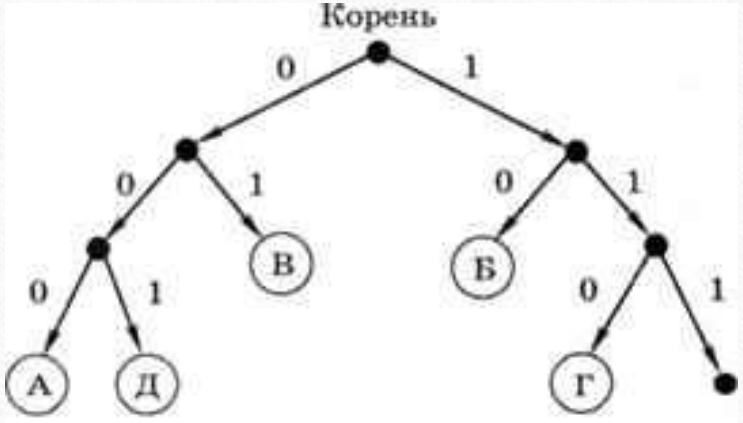
А	Б	В	Г	Д
000	10	01	110	001

- Это неравномерный код, поскольку в нём есть двух- и трёхсимвольные кодовые слова.



Построим для этой кодовой таблицы дерево, в котором от каждого узла (кроме листьев) отходят два ребра, помеченные цифрами 0 и 1.

Чтобы найти код символа, нужно пройти по стрелкам от корня дерева к нужному листу, выписывая метки стрелок, по которым мы переходим (рисунок).



- Заметим, что ни один символ не лежит на пути от корня к другому символу. Это значит, что **условие Фано выполняется** и любую правильную кодовую последовательность можно однозначно декодировать.

Например, рассмотрим цепочку 1100000100110.

1. Букв с кодами 1 и 11 в таблице нет, поэтому сообщение **начинается с буквы Г — она имеет код 110:**

Г	
110	0000100110

2. **Следующий (единственно возможный) код** — 000, это **буква А**:

Г	А	
110	000	0100110

3. Аналогично декодируем всё сообщение:

Г	А	В	Д	Б
110	000	01	001	10