Федеральное государственное образовательное бюджетное учреждение
высшего профессионального образования
Поволжский государственный университет телекоммуникаций и информатики

кафедра Теоретических основ радиотехники и связи

Лабораторная работа «Кодирование дискретных источников»

Руководство пользователя

доц. каф. ТОРС, к.т.н. Чингаева А. М.

Самара, 2012

Содержание

1 Введение

Данная лабораторная работа предназначена для изучения основных методов кодирования дискретных источников. В качестве исходного материала в работе используются тексты на естественных языках. В программе реализованы примитивный равномерный код и разные виды экономных кодов.

2 Интерфейс пользователя

На рис. 1 показано основное окно программы после её запуска.

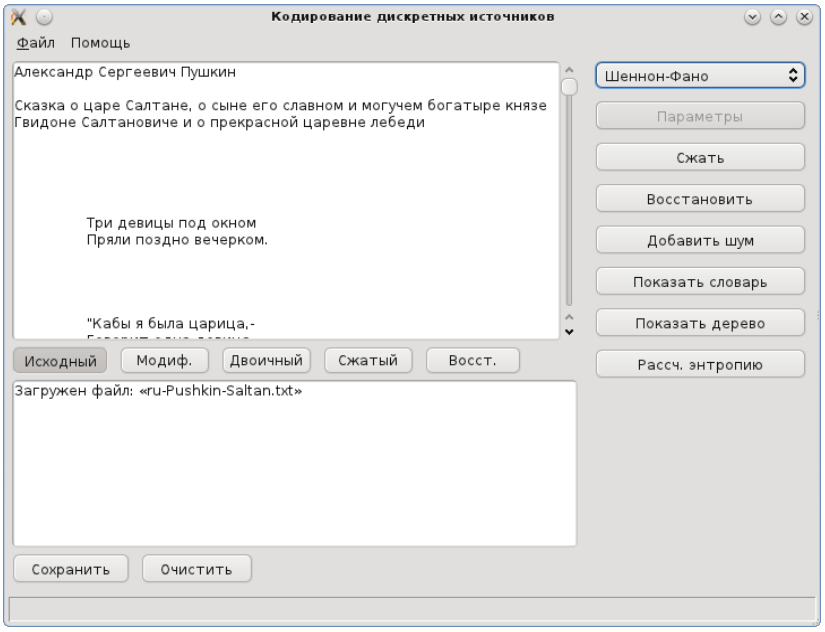


Рис. 1. Окно программы

Элементы управления включают главное меню (см. п. 2.1), выпадающее меню выбора алгоритма сжатия и кнопки действий:

- **Параметры** — изменение параметров кодера (см. п. 2.2).
- **Сжать** — сжимает модифицированный текст (см. п. 2.3) в соответствии с выбранным алгоритмом (см. п. 2.5).
- **Восстановить** — восстанавливает модифицированный текст по сжатому (см. п. 2.5).
- **Добавить шум** — добавляет шум к сжатому двоичному тексту (см. п. 2.6).
- **Показать словарь** — открывает окно обзора словаря (см. п. 2.5).
- **Показать дерево** — отображает кодовое дерево (см. п. 2.5).

- **Рассч. энтропию** — рассчитывает энтропию, используя в качестве исходных данных текущее отображение (см. п. 2.7).

Выпадающее меню позволяет выбирать алгоритм экономного кодирования (сжатия).

Кнопки под окном исходного текста изменяют режим отображения:

- **Исходный** — исходный текст.
- **Модиф.** — модифицированный текст в 32-символьном представлении (см. п. 2.3).
- **Двоичный** — представление модифицированного текста примитивным равномерным двоичным 5-битным кодом (см. п. 2.4).
- **Сжатый** — сжатый текст в двоичном представлении.
- **Восст.** — восстановленный текст в 32-символьном представлении.

При наведении указателя мыши на текст в режимах «Двоичный» и «Сжатый» (для кодов семейства LZ*) отображается всплывающая подсказка:

- буква, соответствующая подсвеченному двоичному коду;
- кодовое слово в десятичном представлении и соответствующий несжатый текст (если это возможно).

Кнопка «Очистить» под окном журнала очищает окно журнала, а кнопка «Сохранить» позволяет сохранить результаты расчётов из журнала в текстовый файл.

2.1 Главное меню

- **Файл**
  - **Открыть** (`Ctrl+O`): загрузить исходный текст из текстового (txt) файла. Unicode–версия программы работает только с текстами в кодировке UTF-8.
  - **Выход** (`Alt+F4`): выйти из программы.
- **Помощь**
  - **Руководство пользователя** (`F1`): открывает это руководство.
  - **О программе**: показывает информацию о программе.

2.2 Параметры кодера

Коды Шеннона-Фано и Хаффмана не имеют параметров, поэтому для них кнопка «Параметры» неактивна.

Для кодов семейства LZ* нажатие кнопки «Параметры» вызывает диалоговое окно, в котором можно изменять

- размер словаря (для всех кодов LZ*),
- максимальную длину фрагмента (только для LZ77).

Оба параметра являются степенью 2 и определяют длину кодового слова соответствующего кода LZ*.

2.3 Модификация исходного текста

Для упрощения представления с целю обучения исходный текст модифицируется следующим образом: все заглавные буквы заменяются на соответствующие строчные, буква «ё» заменяется на «е», а «ъ» на «ь», знаки препинания заменяются пробелами, два и более пробела, следующих подряд друг за другом, заменяются одним. Все прочие символы из текста удаляются. Т.о. для русского языка получается 32-символьная модель: 31 буква плюс пробел.

Все дальнейшие операции проводятся только над модифицированным текстом.

2.4 Примитивное равномерное кодирование

Для примитивного кодирования модифицированного текста используется 5-битный равномерный двоичный код ($n = \log_2 32 = 5$). Результаты примитивного кодирования используются для двоичного отображения модифицированного текста и для расчёта коэффициента сжатия при экономном кодировании.

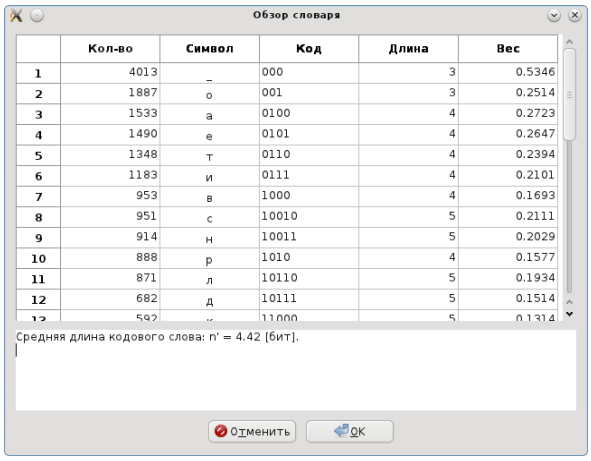
2.5 Экономное кодирование

В работе реализованы 6 различных алгоритмов экономного кодирования (сжатия информации): коды Шеннона-Фано и Хаффмана и три модификации алгоритма Лемпела-Зива — LZ77, LZ78 и алгоритм Лемпела-Зива-Уэлша. Описание работы этих алгоритмов можно найти в соответствующей литературе, например, [1-4].

При сжатии для всех кодов (кроме LZ77) составляется словарь, который можно просмотреть нажав кнопку «Показать словарь».

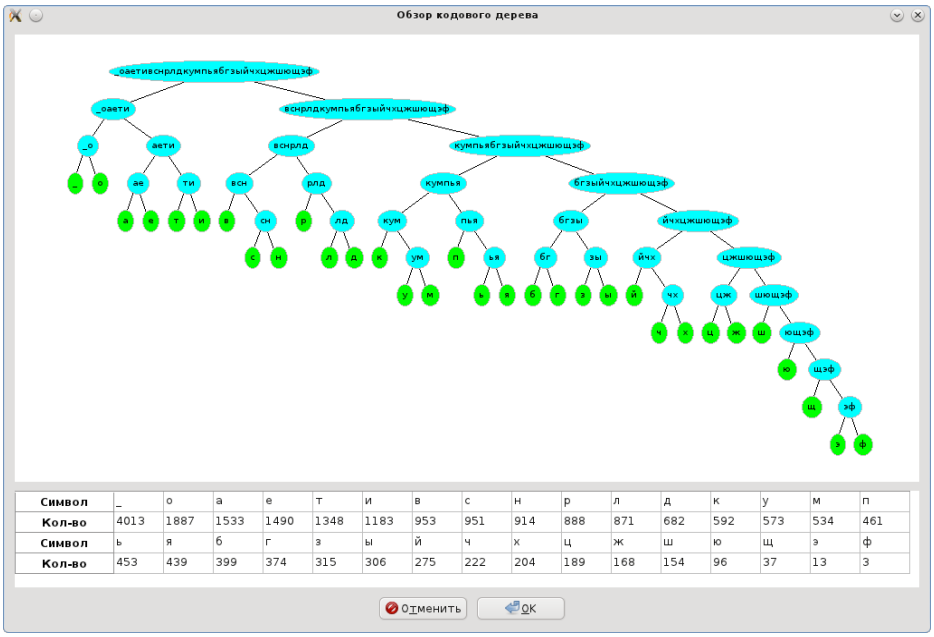
Примерный вид окна обзора словаря показан на рис. 2.

Для кодов Шеннона-Фано и Хаффмана составляется также кодовое дерево, доступное для просмотра при нажатии кнопки «Показать дерево».



	Кол-во	Символ	Код	Длина	Вес
1	4013	_	000	3	0.5346
2	1887	о	001	3	0.2514
3	1533	а	0100	4	0.2723
4	1490	е	0101	4	0.2647
5	1348	т	0110	4	0.2394
6	1183	и	0111	4	0.2101
7	953	в	1000	4	0.1693
8	951	с	10010	5	0.2111
9	914	н	10011	5	0.2029
10	888	р	1010	4	0.1577
11	871	л	10110	5	0.1934
12	682	д	10111	5	0.1514
13	592	к	11000	5	0.1314

Рис. 2. Обзор словаря



Символ	_	о	а	е	т	и	в	с	н	р	л	д	к	у	м	п
Кол-во	4013	1887	1533	1490	1348	1183	953	951	914	888	871	682	592	573	534	461
Символ	ь	я	б	г	з	ы	й	ч	х	ц	ж	ш	ю	щ	э	ф
Кол-во	453	439	399	374	315	306	275	222	204	189	168	154	96	37	13	3

Рис. 3. Кодовое дерево

Примерный вид кодового дерева показан на рис. 3.

Средняя длина кодового слова $n' = \overline{n}$ для кодов Шеннона-Фано и Хаффмана рассчитывается как средневзвешенное значение n_i (длин отдельных кодовых комбинаций)

$$\overline{n} = \sum_{i=1}^{K} n_i P(a_i).$$

Для кодов семейства LZ* величина $n' = \overline{n} = n$ является фиксированной и определяется размером словаря и видом кода:

$n = \log_2 D + \log_2 L + 5$ — для кода LZ77,

$n = \log_2 D + 5$ — для кода LZ78,

$n = \log_2 D$ — для кода LZW.

Где D — размер словаря, L — максимально допустимая длина фрагмента, 5 — количество бит для представления исходного символа источника.

Также для всех кодов при сжатии рассчитывается коэффициент сжатия:

$$r = \frac{N_{\text{сж}}}{N_{\text{мод}}} \cdot 100\%.$$

Здесь $N_{\text{сж}}$ — длина сжатого текста в битах, $N_{\text{мод}}$ — длина модифицированного текста, закодированного двоичным примитивным кодом.

2.6 Шум

Кнопка «Добавить шум» позволяет изучить влияние ошибок на восстановление сжатого текста. Однократное нажатие кнопки добавляет к сжатому двоичному тексту двоичный шум с $P(\text{ош}) = 0{,}01$.

2.7 Расчёт энтропии

Расчёт энтропии производится для посимвольной модели без учёта связей между соседними символами.

3 Литература

1. Кловский Д. Д. Теория электрической связи. — М.: Радиотехника, 2009. — 648 с.

2. Теория электрической связи: учебник для вузов / А. Г. Зюко, Д. Д. Кловский, В. И. Коржик, М. В. Назаров; Под ред. Д. Д. Кловского. — М.: Радио и связь, 1998. — 432 с.

3. Шеннон К. Работы по теории информации и кибернетике. — М,: Издательство иностранной литературы, 1963.

4. Яглом И. М., Яглом А. М. Вероятность и информация. — М.: Наука, 1973.