Символы

Д. В. Луцив

Кафедра системого программирования СПбГУ

CS-103

Содержание

1. Ъ
   - Значение символа
2. Алфавит и кодировка
   - Устная речь и переписывание
   - До- и раннекомпьютерное кодирование
   - Компьютерное кодирование
3. Unicode
   - Основы
   - Пространство
   - Нормализаиця
   - Порядок
   - Кодировки

- Луців — і \u0456 кириллическая десятиричная і. Не то, что латинская.
- В русском языке — «Абхазія», «Авіаторъ», «Кіевъ».
- В украинском — «місто». «Київ» — не как в русском.

і

- В шрифтах хранится как *одна* картинка.
- Представление с латиницей разное,
- но общее среди языков кириллицы.
- Семантика разная в русском украинском.

И

- Есть в языках кириллицы.
- Изображение и символ одни и те же.
- Семантика в русском и украинском разная.

Ё

В современном русском только 2 буквы с диакритикой - й и ё.
Ё ввели (по разным версиям):

- Княгиня Екатерина Дашкова (директор РАН) 1783 г.
- Николай Карамзин в слове «сліозы», 1797 г.

- При устной речи символов, как таковых, нет вообще.
- При переписывании в ручную и передаче по фототелеграфу проблем нет, т.к. есть начертание и семантика, но нет кодировки. Есть форматирование и диакритика. И даже иллюстрации иногда сами вставляются.

Историческая справка 1

Параллельное и смешанное

- Семафорный телеграф — братья Шапп, 1780, 196 положений, 2 сл./мин.
- Ящик со шрифтом, типография.
- Современный телеграф.

Последовательное

- Оптический (и солнечный) телеграф.
- Код Морзе.
- Тюремная азбука.

Историческая справка 2

- Викторианский телеграф.
- Код Бодо, 5 битов.
- МТК-1 и МТК-2.

Ранние

- Первые — телеграфные, и байт из 6 битов.
- Стандартные — до сих пор ASCII и EBCDIC.

Современные

- Производные от ASCII.
- Unicode.

Мотивы прогресса

- обработка максимум двуязычных текстов, один из них английский
- перекодировки

Пространство

$$2^{16} + 2^{20} = 2^{16} + 2^4 \times 2^{16} = 17 \times 2^{16}$$

Символы

- протяжённые (обычные самостоятельные)
- непротяжённые
  - управляющие (право-левые, переворот, т.д.)
  - диакритические
  - модификаторы начертания

$$2^{16} + 2^{20} = 2^{16} + 2^4 \times 2^{16} = 17 \times 2^{16}$$

- 2^{16} — `U+XXXX` — базовая многоязычная плоскость
- $2^4 \times 2^{16}$ — `U+1XXXX`...`U+10XXXX` — остальные.

- Канонический класс модификации — неотрицательное число, поставленное в соответствие символу. Для обычных символов 0, для диакритических знаков м.б., например, 230.
- Канонический порядок — выстраивание с консервативной (устойчивой) сортировкой по возрастанию класса модификации.
- Эквивалентность
  - Каноническая — если эквивалентны канонические декомпозиции ế (U+1EBF) (вьет.) — м.б. и неоднозначно, т.к. у диакритик акута и циркумфлекса один класс.
  - Совместимости — например, нижний индекс «$_2$» и «2» — которая может быть отнесена к форматированию.

II

- (NFD) Каноническая декомпозиция — преобразование готовых символов с модификаторами в протяжённый символ и последовательность модификаторов в каноническом порядке. Например, Ç → C[25CC][327].
- (NFC) Каноническая декомпозиция с последующей композицией (см. конспект).
- KD (NFKD) — совместимая декомпозиция. При приведении в эту форму все составные символы заменяются используя как канонические карты декомпозиции Юникода, так и совместимые карты декомпозиции, после чего результат ставится в каноническом порядке.
- KC (NFKC) — совместимая декомпозиция с последующей канонической композицией.

Не соответствуют кодам вообще. Отношение порядка — атрибут не набора символов и даже не письменности, а языка. Сравнение многоуровневое (английский язык):

1. Буквы — role < roles < rule
2. Диакритика — role < rôle < roles
3. Регистр — role < Role < rôle
4. Пунктуация — role < “role” < Role
5. Форматирование и управление — role < r**o**le < “role”

1. Для сравнения канонически эквивалентных строк модификаторы одинакового класса можно тоже отсортировать.
2. Таблица порядка — таблица для языка, в которой перечислены веса символов как минимум до уровня 3. Если нет частного случая для конкретного символа, можно его декомпозировать.
3. Всякие исключения:
   - Английский — cote < coté < côte < côté
   - Французский — cote < côte < coté < côté у последней диакритики большее значение

- UTF-32 (UCS-4) — постоянная длина, но по 4 байта на символ много.
- UTF-16 (UCS-2)
  - постоянная длина для нулевой плоскости.
  - Встроенная для Java, .NET и многих дрругих современных платформ.
  - Для перехода на другие плоскости — упр. символы. Символы с кодами меньше 0x10000 (2^{16}) как есть, символы с кодами `0x10000-0x10FFFE` — 2 слова, `0xD800-0xDBFF` и `0xDC00-0xDFFF`. $2^{10} \times 2^{10} = 2^{20}$ таких комбинаций
  - бывает LE и BE.
- UTF-8 - переменной длины.

UTF-8

Unicode	Byte1	Byte2	Byte3	Byte4
U+000000-U+00007F 0xxxxxxx	0xxxxxxx			
U+000080-U+0007FF 00000yyy xxxxxxxx	110yyyxx	10xxxxxx		
U+000800-U+00FFFF yyyyyyyy xxxxxxxx например Евро	1110yyyy	10yyyyxx	10xxxxxx	
U+010000-U+10FFFF 000zzzzz yyyyyyyy xxxxxxxx (не BMP)	11110zzz	10zzyyyy	10yyyyxx	10xxxxxx

- Punycode - для URL, с сохранением ASCII-фрагмента. http://1ш2щ3.ru/ $\rightarrow$ http://www.xn--123-xfdg.ru/
  - Забавный алгоритм; разобраться самим

Вопросы

EDU.DLUCIV.NAME