МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ЯЗЫК И КУЛЬТУРА

Сборник статей
XXX Международной научной конференции
(16–19 сентября 2019 г.)

Ответственный редактор
доктор педагогических наук, профессор *С.К. Гураль*

Томск
Издательский Дом Томского государственного университета
2020

Е.В. Трусов

Национальный исследовательский Томский государственный университет

Сравнение систем кодировки английского и китайского языков
Comparing of English and Chinese encoding systems

Аннотация. В статье рассматриваются различия систем представления текстовой информации в памяти компьютера, сходства и отличия кодирования иероглифов и букв английского алфавита.
Abstract. This article explains differences of representation of text information in computer memory, similarities and differences of coding characters and letters of English alphabet.

В веке информационных технологий многие люди, изучая иностранные языки, не знают, как символы, используемые ими в повседневной жизни и в процессе обучения, записываются в памяти устройств, которыми они пользуются. Этот вопрос является актуальным, потому что у студентов должна быть базовая информация об используемых системах кодирования компьютером для прояснения отдельных вопросов, например, отсутствия какого-то конкретного иероглифа в системе кодирования. Статья будет наиболее полезна тем, кто заинтересован во взаимосвязи информационных технологий с лингвистикой и филологией. Данная работа описывает использование систем кодировки, записи символов, в частности иероглифов китайского языка и букв английского алфавита, а также определяет сходства и различия в кодировании этих символов компьютером.

Под системой кодировки понимается совокупность правил представления символов в памяти компьютера. Система кодирования представляет собой правило замены одного объекта на код. Соответственно, код – условное обозначение объекта знаком или группой знаков по определенным правилам, установленным системами кодирования.

Первые системы кодирования символов появились еще в XIX веке. Первооткрывателем всех систем кодировок принято считать Сэмюэла Морзе, который в 1837 году изобрёл одноименную азбуку («Азбука Морзе»). Данная азбука использовалась в телеграфном сообщении, которое является последовательностью электрических сигналов, пе-

редаваемых от одного телеграфного аппарата по проводам к другому телеграфному аппарату.

С появлением новых технологий передачи информации появилась необходимость в создании особых систем записи данных в памяти компьютера. В 1963 году в Америке была разработана первая система кодирования символов для компьютера, которая называется ASCII 7. Эта система содержала в себе только буквы английского алфавита, десятичные цифры, знаки препинания, управляющие символы.

ASCII 7-система кодировки, представленная в виде таблицы и содержащая в себе заглавные и строчные буквы латинского алфавита, цифры, специальные символы и знаки препинания. В этой системе кодировки есть 128 символов, которые записаны от 0 до 127 в двоичной системе счисления в памяти компьютера. Заглавные символы латинского алфавита записаны с 65 по 90 номер в таблице, а строчные – с 97 по 122. Это сделано для упрощения таблицы и ускорения быстродействия компьютера в наборе и поиске необходимого символа для его кодирования и выведения на экран. Рассмотрим сказанное выше на конкретном примере.

Символ «А» в десятичной системе счисления записан как 65, в двоичной – 01000001. Буква «а» в двоичной системе счисления-01100001, а в десятичной, соответственно, 97.

Приведенный выше пример показывает, что заглавные и строчные символы, представленные в кодировке ASCII 7, отличаются друг от друга на 32 и на 00100000 в двоичной системе счисления, что облегчает работу процессору для совершения операции поиска необходимого символа в таблице кодирования.

Для кодирования текстов, написанных на английском языке, ASCII 7 было вполне достаточно, однако что же делать, если текст нужно представить на другом языке или на нескольких языках одновременно? Ответ на данный вопрос – появление уникальной системы под названием «Unicode». Unicode – стандарт кодирования символов, который включает в себя знаки почти всех письменных языков мира.

GB18030 – кодировка, утвержденная правительством Китайской Народной Республики, основанная на принципе системы Unicode. GB18030 содержит в себе 6737 китайских символа, пиньин, японскую кану, чжуинь, кириллицу, символы пиньина с диакритикой и символы пунктуации. Путём несложных математических вычислений выводим, что данная система кодировки имеет возможность закодировать око-

ло 45 млн символов, чего достаточно, чтобы содержать в себе 99,75% всех существующих китайских иероглифов, включая традиционные, такой процент получается в результате использования китайцами уникальных иероглифов для имен собственных.

Принцип кодирования символов в GB18030 основан на одновременном использовании 2 байт информации совместно с CJK Unified Ideographs (использующей 4 байта памяти). Формат разделения 2 байт памяти, отведённых под 1 иероглиф, на 16 минимальных ячеек памяти компьютера (битов), позволяет точно передать любой иероглиф.

Рассмотрим отличия систем кодировок ASCII и GB18030. ASCII представляет собой таблицу, в которой все символы упорядочены и имеют логическую связь, в то время как GB18030 представляет собой таблицу составных элементов и ключей, из которых состоят иероглифы. Для записи символов английского алфавита используется 1 байт информации, в то время как для записи иероглифа при помощи таблицы кодирования GB18030 используется 2 байта информации, что делает текст, содержащий одинаковую информацию на разных языках, автоматически в 2 раза большим по объёму памяти. Методы кодирования символов во многом отличаются друг от друга. При кодировании текстовых данных, используя систему ASCII 7, байт информации не разделяется на отдельные его составные части, чего нельзя сказать о системе GB18030, где 2 байта информации составляются из 16 битов. Существует колоссальное отличие в количестве самих символов, содержащихся в системах, в GB18030 в несколько раз больше символов, чем в ASCII 7, что, конечно же, обуславливается спецификой китайского языка.

Однако для кодирования текста на разных языках мира был создан большой массив данных под названием UTF 32, основанный на принципе работы системы кодирования Unicode, который описан ранее. Главная особенность utf 32 заключается в прямой индексации символов, что делает обработку каждого символа одинаковой по времени, длина кода при использовании utf 32 также не изменяется. UTF 32 использует одновременно 4 байта информации, что делает его универсальным в использовании одновременно для большинства языков, представленных в системе windows.

Литература

1. Standard ECMA-6: 7-bit coded character set. 6th edition Ecma international December 1991.

2. Tom Jennings. An annotated history of some character codes or ASCII. American standard code for information infiltration (1999-09-16 – 2004-10-29)
3. Anthony Fok. "Application of IANA Charset Registration for GB18030". IANA Character Set Registrations. Retrieved 2002-03-15 – 2016-12-05.
4. Standardization Administration of China (SAC) (2005-11-18). GB 18030-2005: Information Technology–Chinese coded character set.
5. David C. Zentgraf. What Every Programmer Absolutely, Positively Needs To Know About Encodings And Character Sets To Work With Text.
6. https://unicode-table.com/ru/alphabets/chinese/
7. https://unicode-table.com/ru/blocks/cjk-unified-ideographs/