Классная работа

Инструменты для распознавания текстов и системы компьютерного перевода. Оценка количественных параметров текстовых документов

Программы оптического распознавания документов

Для ввода текстов в память компьютера с бумажных носителей используют **сканеры** и **программы распознавания символов**.

Одной из наиболее известных программ такого типа является **ABBYY FineReader**.

Программы оптического распознавания документов

Вместо сканера можно использовать цифровой фотоаппарат или камеру мобильного телефона.

Программа ABBYY FineReader

Одна из лучших в мире программа для оптического распознавания текста (192 языка). Разработана для операционных систем Microsoft Windows, macOS и Linux (проприетарное программное обеспечение).

Программа позволяет **сканировать** и преобразовывать с **оптическим распознаванием** изображения документов (фотографий, результатов сканирования, PDF-файлов) в электронные редактируемые форматы:

- Microsoft Word
- Microsoft Excel
- Microsoft Powerpoint
- RTF
- HTML
- PDF
- текстовый файл

Программа распознаёт множество языков, в том числе в одном тексте.

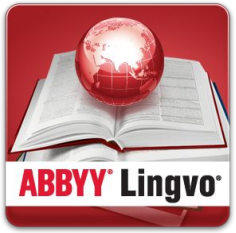
Компьютерные словари

Компьютерные словари выполняют перевод отдельных слов и словосочетаний.

Компьютерные словари обеспечивают мгновенный поиск словарных статей.

Многие словари предоставляют пользователям возможность прослушивания слов в исполнении носителей языка.

Программа ABBYY Lingvo



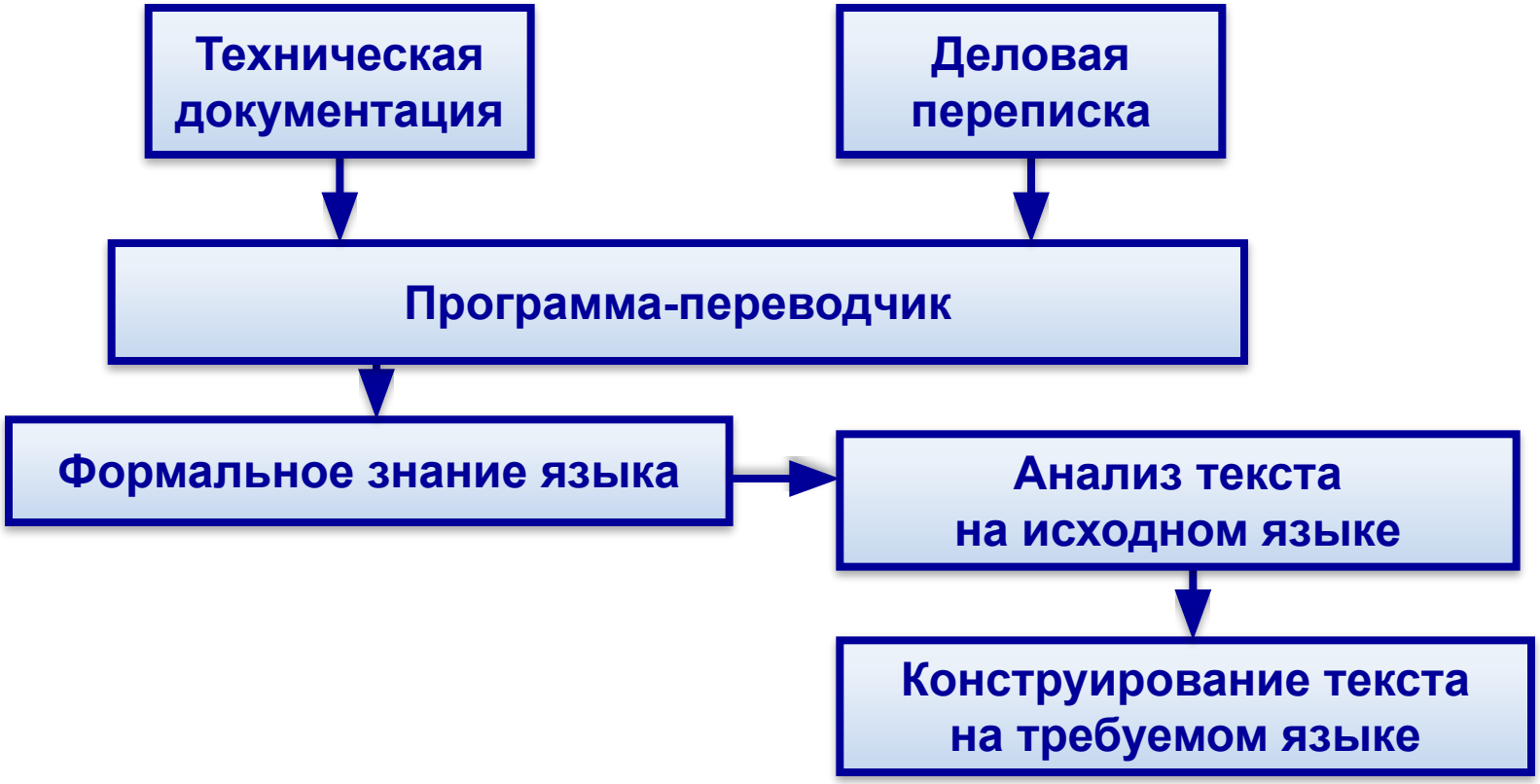
Одной из наиболее известных программ-словарей. Имеются пакеты для многих популярных операционных систем, таких как Windows, Windows Mobile, Symbian OS, Mac OS X, iOS, Android (проприетарное программное обеспечение).

Включает сотни общелексических и тематических словарей для **перевода** и **толковых** словарей.

В ABBYY Lingvo **нет** функции полнотекстового перевода, но возможен пословный перевод текстов из буфера обмена.

Компьютерные программы-переводчики

Для перевода текстовых документов применяются **программы-переводчики**.



Художественный литературный перевод текста пока не возможен.

Программа PROMT

Одной из наиболее известных у нас программ-переводчиков является **PROMT** от одноимённой российской компании **PROMT** (проприетарное программное обеспечение).

Существует множество различных пакетов для дома и бизнеса. Множество языков.

Google Переводчик

Google Переводчик (англ. **Google Translate**) — веб-служба компании Google, предназначенная для автоматического перевода части текста или веб-страницы на другой язык.

translate.google.com

На сегодня в переводчике доступны 103 языка.

Представление текстовой информации в памяти компьютера

Текст состоит из символов - букв, цифр, знаков препинания и т.д., которые компьютер различает по их **двоичному коду**.

- на экране (символы)

- в памяти – двоичные коды

01000001	01000010	01000011	01000100
65	66	67	68

Соответствие между изображениями символов и кодами символов устанавливается с помощью **кодовых таблиц**.

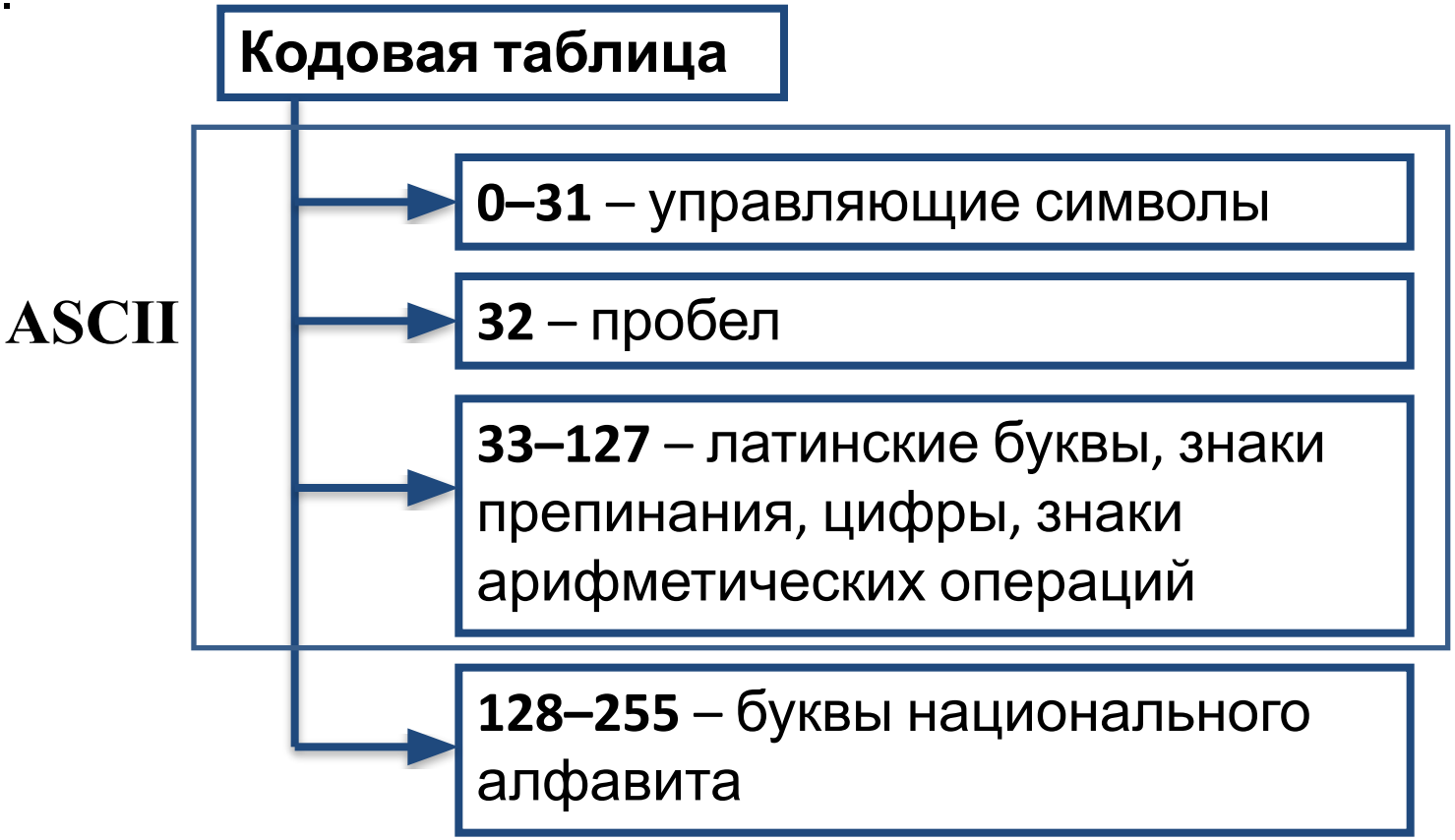
Фрагмент кодовой таблицы ASCII

Символ	Десятичный код	Двоичный код	Символ	Десятичный код	Двоичный код
Пробел	32	00100000	**0**	48	00110000
!	33	00100001	**1**	49	00110001
#	35	00100011	**2**	50	00110010
$	36	00100100	**3**	51	00110011
*	42	00101010	**4**	52	00110100
=	43	00101011	**5**	53	00110101
,	44	00101100	**6**	54	00110110
-	45	00101101	**7**	55	00110111
_	46	00101110	**8**	56	00111000
/	47	00101111	**9**	57	00111001
A	65	01000001	**N**	78	01001110
B	66	01000010	**O**	79	01001111
C	67	01000011	**P**	80	01010000

и

8-битные коды русских букв

Код ASCII содержит изначально 128 символов (0–127). Среди них нет русских букв. 8-битные коды имеют 256 кодовых комбинаций ($2^8 = 256$). Коды от **128 до 255 использовали для кодирования** букв национального алфавита (в разных странах по разному, в одной стране оказалось несколько разных кодовых таблиц).



Коды русских букв в разных кодовых таблицах

Символ	Кодовые таблицы			
	Windows		КОИ-8	
	десятичный код	двоичный код	десятичный код	двоичный код
А	192	11000000	225	11100001
Б	193	11000001	226	11100010
В	194	11000010	247	11110111

Увеличение мощности кода

Кодовая таблица символов **Unicode** позволяет пользоваться более чем двумя языками в одном тексте.

В **Unicode** каждый символ кодируется **шестнадцатиразрядным** двоичным кодом. Такое количество разрядов позволяет закодировать **65 536** различных символов: $2^{16} = 65\,536$.

При использовании кодовой таблицы **Unicode** в тексте одновременно могут содержаться любые символы всех языков мира.

Информационный объём фрагмента текста

Информационный объём фрагмента текста – это количество бит, байт (килобайт, мегабайт), необходимых для записи фрагмента оговорённым способом кодирования.

$$I = K \times i$$

I – информационный объём сообщения
K – количество символов
i – информационный вес символа

В зависимости от разрядности используемой кодировки информационный вес символа текста может быть равен:

- 8 бит (1 байт) - **восьмиразрядная кодировка**;
- 16 бит (2 байта) - **шестнадцатиразрядная кодировка**.

Задача 1

Задача 1. Считая, что каждый символ кодируется одним байтом, определите, чему равен информационный объём следующего высказывания Жан-Жака Руссо:

Тысячи путей ведут к заблуждению, к истине – только один.

Решение

В данном тексте 57 символов (с учётом знаков препинания и пробелов). Каждый символ кодируется одним байтом. Следовательно, информационный объём всего текста – 57 байт.

Ответ: 57 байт.

Задача 2

Задача 2. В кодировке Unicode на каждый символ отводится два байта. Определите информационный объём слова из 24 символов в этой кодировке.

Решение.

$I = 24 \times 2 = 48$ (байт).

Ответ: 48 байт.

Задача 3

Задача 3. Автоматическое устройство осуществило перекодировку информационного сообщения на русском языке, первоначально записанного в 8-битовом коде, в 16-битовую кодировку **Unicode**. При этом информационное сообщение увеличилось на 2048 байтов. Каков был информационный объём сообщения до перекодировки?

Решение

Информационный вес каждого символа в 16-битовой кодировке в два раза больше информационного веса символа в 8-битовой кодировке. Поэтому при перекодировании исходного блока информации из 8-битовой кодировки в 16-битовую его информационный объём должен был увеличиться вдвое, другими словами, на величину, равную исходному информационному объёму. Следовательно, информационный объём сообщения до перекодировки составлял 2048 байтов = 2 Кб.

Ответ: 2 Кбайта.

Задача 4

Задача 4. Выразите в мегабайтах объём текстовой информации в «Современном словаре иностранных слов» из 740 страниц, если на одной странице размещается в среднем 60 строк по 80 символов (включая пробелы). Считайте, что при записи использовался алфавит мощностью 256 символов.

Решение

$K = 740 \times 80 \times 60$ $N = 256$	$I = K \times i$ $N = 2^i$
I - ?	

$256 = 2^i = 2^8$, $i = 8$

К = 740 × 80 × 60 × 8 = 28 416 000 бит = 3 552 000 байт =

= 3 468,75 Кбайт $\approx$ 3,39 Мбайт.

Ответ: 3,39 Мбайт.

Задание

- Откройте стр. **193** – Задание **4.16**.
- Создайте в **личной папке** (папка **Фамилия**) файл типа **документ Word** с именем **Формулы**.
- Выполните задание **4.16**.
- Закройте файл **с сохранением**.

Работаем за компьютером

Домашнее задание

Прочитать **§4.5** (стр. 174–177). Задания **2–6** (стр. 177) – **устно**. Задание **3** (стр. 177) – **письменно**.
Выучить **§ 4.6** (стр. 178–183). Повторить **§ 1.6** (стр. 45–50). Задания **2-7** (стр. 183–184) – **устно**.
Задания **8** (стр. 184) – **письменно**.
Написать конспект по презентации, выучить, выполнить задания.
Готовый файл пришлите мне на электронную почту akiwina82@mail.ru