**ФГБОУ ВПО «Воронежский государственный университет»**

**Факультет романо-германской филологии**

**КУРСОВАЯ РАБОТА**

**«ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ОБРАБОТКЕ ТЕКСТОВ»**

**Исполнитель:**

**Студентка 1 курса**

**Воробьева Екатерина Владимировна**

**Проверила:**

**Донина О.В.**

**Воронеж 2018**

Оглавление

[Вступление 3](#_Toc526168684)

[1. Текст: представление, хранение, ввод 4](#_Toc526168685)

[Представление текста 4](#_Toc526168686)

[Текстовые файлы 5](#_Toc526168687)

[Правила машинописного набора текста 6](#_Toc526168688)

[2. Оформление текста 7](#_Toc526168689)

[Шрифты 7](#_Toc526168690)

[Структурирование теста 9](#_Toc526168691)

[3. Автоматизированная обработка текста 11](#_Toc526168692)

[Поиск 11](#_Toc526168693)

[Расшифровка или уточнение значений слова 12](#_Toc526168694)

[Заключение 13](#_Toc526168695)

[Список использованной литературы 14](#_Toc526168696)

# Вступление

Традиционно под словом «технология» понимают процедуры и средства, позволяющие получить определенный результат в какой-то области человеческой деятельности. Информационные технологии в этом смысле не являются исключением – под ними понимается вся совокупность программного и аппаратного обеспечения, методов и средств организации автоматизированной обработки информации, представленной в цифровой форме.

Постоянный рост мощности вычислительной техники увеличивает спектр решений, доступных пользователю современного компьютера. Решение многих типовых на сегодняшний день задач еще 10 лет назад представляло серьезную, часто неразрешимую проблему. Многие технологии были не доступны за пределами крупных организаций. Постоянное и быстрое развитие техники и программного обеспечения приблизило их к пользователю. Следует отметить, что одним из магистральных направлений развития современных информационных технологий является разработка подходов и методов интеграции данных различного вида, смешивания функций различных систем, активное использование методов коммуникации для обмена данными.

# 1. Текст: представление, хранение, ввод

## Представление текста

Представление информации в виде текста стало одним из первых доступных для обработки с помощью ЭВМ и до сих пор остается одним из наиболее универсальных. Энциклопедический словарь дает такое определение понятию “текст”: “**Текст** - это*упорядоченный набор слов, предназначенный для того, чтобы выразить некий смысл. В лингвистике термин используется в широком значении, включая в себя и устную речь*”.[[1]](#footnote-1)

Представление информации в виде текста при обработке с помощью вычислительной техники близко к этому определению. Под “текстовым” понимают такое представление информации, в котором она представлена в виде записи слов (логических элементов) некоторого языка и доступна для чтения человеком.

Язык для такого представления характеризуется некоторым алфавитом - т.е. допустимым набором символов. Поскольку компьютер работает только с двоичным кодом, то для записи и обработки требуется взаимно-однозначно сопоставить символы и двоичные коды. *Правило сопоставления кодов и символов, входящих в алфавит, называется* *кодировкой.*

Первый широко распространенный стандарт кодирования - таблица (т.е. прямое сопоставление кодов символам) кодировки ASСII*,* - был разработан в 1963 году. Стандарт предполагал использование не только в вычислительной технике, но и в телеграфии. В нем для кодирования каждого символа отводилось 7 бит. Восьмой бит использовался для служебных целей - контроля четности при передаче.

Эта часть таблицы кодировки содержит символы латинского алфавита, цифры, некоторые знаки препинания и набор управляющих символов (возврат каретки, перевод строки, конец файла, сигнал и т.п.).

Позже восьмой бит стали использовать для представления символов национальных алфавитов: первая часть таблицы - US-ASCII - использовалась по-прежнему, а содержание второй менялось в зависимости от исходного естественного языка. Каждый вариант этой второй половины (расширенной таблицы) исходной таблицы получил название “кодовой страницы” языка (code page).

Для русского языка таких расширений несколько. Наиболее известны: CP866 (DOS), KOI-8R (UNIX), CP1251 (Windows) и MacCyr.

Применение такого способа кодирования сильно затрудняет передачу текстовых сообщений между разными странами, объединение в сообщении текста на нескольких языках, а в случае с русским языком - и обмен файлами между разными ОС (для русского языка до сих пор активно применяется 4 разных кодовых таблицы). Для решения этих проблем в 1991 году некоммерческим объединением был предложен стандарт кодирования Юникод (Unicode).

Стандарт состоит из двух частей: универсального набора символов (Universal Character Set) и правил трансформации (Unicode Transformation Format). Универсальный набор символов предполагает описание всех возможных при записи текстов символов в виде общей таблицы кодов. Правила трансформации определяют способ записи этих кодов.[[2]](#footnote-2)

Первая версия стандарта предполагала использование двух байтов для кодирования каждого символа. В дальнейшем это кодовое пространство было расширено.

Сейчас чаще всего применяется способ трансформации UTF-8, обеспечивающий совместимость с предыдущими реализациями и стандартами. В частности, коды менее 128 записываются одним байтом, что автоматически превращает их в коды ASCII.

Применение этого стандарта кодирования позволяет объединять в одном тексте слова на различных языках (без ограничений на их количество), использовать устаревшие языки, дополнительные символы.

Наиболее переносимым и легко используемым с технической точки зрения способом хранения и передачи текста являются текстовые файлы. По сути, эти файлы представляют собой последовательности символов, разбитых на абзацы или строки.

## Текстовые файлы

Понятие “текстового файла” не предусматривает строго заданного формата или расширения. Тем не менее, помимо характерной для той или иной ОС таблицы кодировки, в текстовых файлах могут применяться три основных способа деления текста на абзацы (табл.1).

Таблица 1

|  |  |
| --- | --- |
| Windows (DOS) | символы “Возврат каретки” + “Перевод строки” (CR+LF) |
| Unix | символ “Перевод строки” (LF) |
| MacOs | символ “Возврат каретки” (CR) |

Текстовые файлы применяются для самых различных целей и часто оказываются формой хранения данных, описанных более сложными формальными языками. Эти файлы часто используются для записи конфигурации ПО, документирования, переноса данных, описания HTML- или XML-кода.

## Правила машинописного набора текста

Для облегчения анализа и последующего преобразования текста при его наборе в самых различных случаях рекомендуется соблюдать общие правила машинописного набора:

1. Все слова разделяются пробелом, и только одним пробелом.

2. Знаки препинания примыкают к предыдущему слову.

3. Скобки и кавычки всех видов примыкают к первому и последнему слову заключенного в них текста.

4. Текст разрывается только в конце абзаца.

5. Большие форматированные пробелы делаются вставкой символа табуляции, а не несколькими пробелами подряд.

Соблюдение этих правил позволяет легко использовать текст при подготовке более сложных документов, в которые он входит как важнейший элемент, или при организации автоматической обработки.

Текст может появиться из самых разных источников. Чаще всего текстовую информацию вводят с помощью клавиатуры. Стандартная клавиатура и программа, принимающая от нее информацию о нажатых клавишах, позволяют вводить текст, указывать место ввода в уже введенном тексте (перемещая маркер места ввода клавишами перемещения курсора либо с помощью мыши) и удалять неверно введенные символы слева или справа от курсора.

Возможность исправлять ошибки и набирать текст постепенно стала одной из существенных причин, по которым подготовка текстовой информации практически повсеместно была переведена с бумажной на компьютерную основу.

Текстовые редакторы с развитыми возможностями предоставляют пользователям возможность протоколировать и сохранять наборы действий - создавать макрокоманды, или **макросы**. Использование макросов позволяет ускорить выполнение частых простых задач обработки.

Специализированные программы, основной задачей которых является обеспечение набора текста, разделяют на **текстовые редакторы**

(т.е. *программы, которые помогают именно подготовить тот или иной специфический текст, но не оформить его для печати*) и **текстовые процессоры (**т.е.*более сложные* *программные комплексы, позволяющие выполнить оформление текста, точно задать его расположение, сопроводить его графическими материалами)*.[[3]](#footnote-3)

**Пример программных продуктов текстовых редакторов:**

Блокнот, Notepad++, PSPad, vi

# 2. Оформление текста

## Шрифты

Чаще всего текстовая информация используется при подготовке различных печатных материалов. Конечная цель подготовки такого материала - его печать или точное изображение печатной страницы на экране. В отличие от простой подготовки текстового файла, при подготовке печатного материала важно, как отображается текст. Практически все основные элементы и приемы оформления текстовых материалов заимствованы у давно существующих технологий - печатных, оттуда же пришла и основная часть терминологии.

Основным и наиболее важным средством определения внешнего вида текста является **шрифт**(*schreiben*, от нем. - “писать”). Шрифт - это графический рисунок букв, цифр и символов, обладающий общими для всех символов стилистическими особенностями изображения.

Отдельный символ контурного шрифта с обозначенными элементами рисунка.

Шрифт характеризуется рядом параметров:

1. Рисунок шрифта - графические особенности, определяющие общность шрифта и его отличие от всех других.

2. Кегль - размер шрифта - предельная высота большой буквы и окружающих ее пробелов (термин введен для описания высоты площадки литеры при наборе с помощью типографской кассы). Чаще всего задается в типографских пунктах (1 пункт = 1/72 дюйма = 0,375 мм). По историческим причинам некоторые размеры имеют собственные названия: 8 пт - “петит”, 9 пт - “боргес”, 10 пт - “корпус”, 12 пт - “цицеро”.

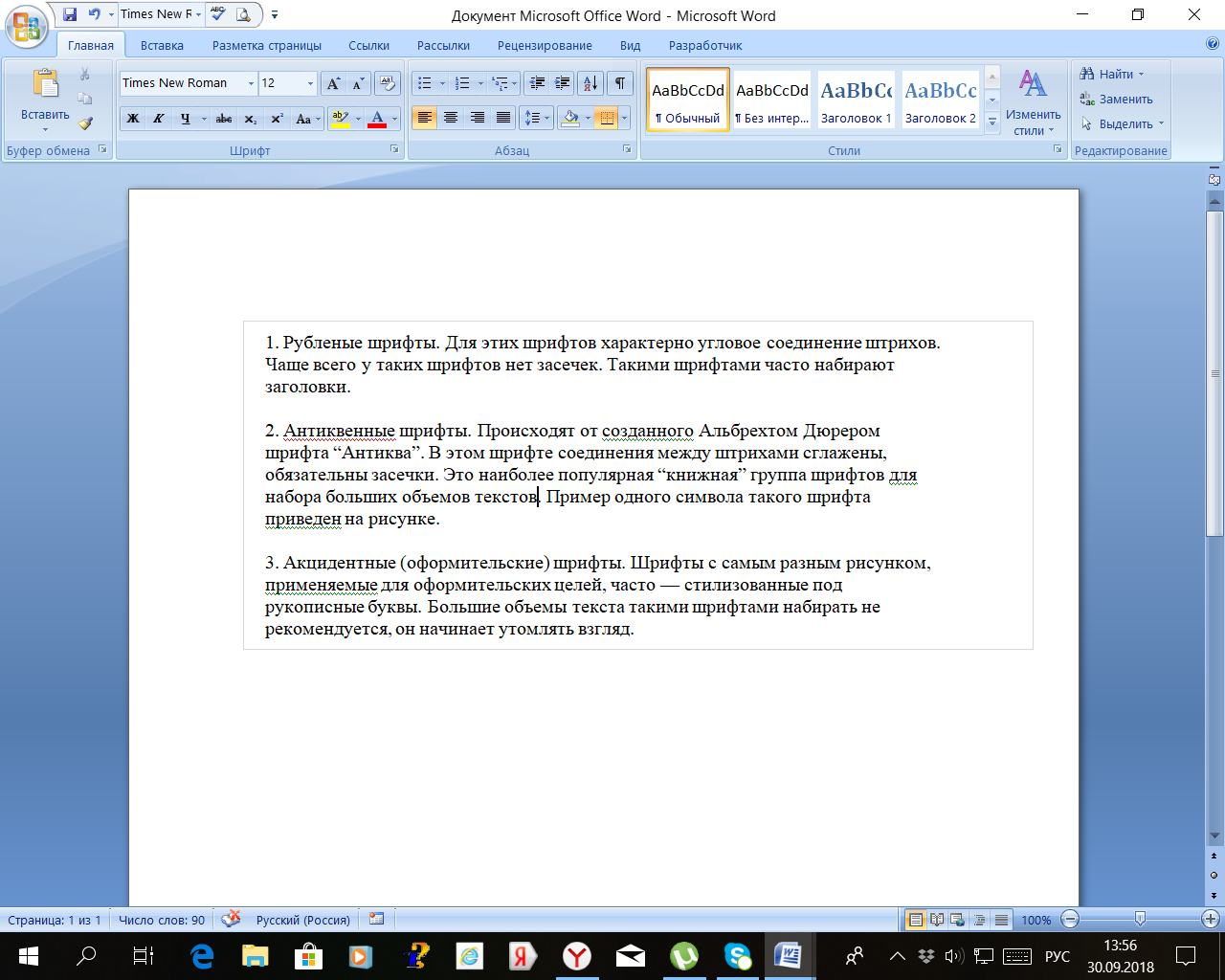
3. Начертание - шрифт с общим рисунком, но какими-либо отличительными признаками: более жирный, наклонный, разреженный. Иногда параметр плотности шрифта (светлый, полужирный, жирный) отделяют от начертания.

4. Часто как параметр задается подчеркивание или зачеркивание шрифта, или его написание как индекса - с уменьшением размера и подъемом/спуском относительно текущей строки.

Совокупность всех возможных размеров и вариантов написания шрифта называется **гарнитурой**. Гарнитуры имеют имена, по которым часто называют и конкретный шрифт.

По общим чертам рисунка различают три основных вида шрифтов (рис.1).

Рис.1



Шрифт задается для набранного текста и не изменяет самих символов - он только определяет написание каждого символа, исходя из эталонного изображения. Библиотека таких изображений называется просто “шрифтом”.

Существует несколько основных способов описания шрифтов (точнее - гарнитуры шрифта):

*Растровые шрифты.* При таком способе каждая буква описывается отдельно, как некоторая матрица точек. Способ позволяет максимально ускорить обработку, но сильно затрудняет изменение размеров или начертаний. Для достижения качества каждый символ такой гарнитуры должен быть отредактирован вручную и должен храниться отдельно.

*Векторные шрифты.* При таком способе описания шрифт задается с помощью некоторых математических кривых, совокупность которых и составляет рисунок символов. Такой шрифт может изменять размеры без потери качества, но с помощью примитивов трудно добиться прорисовывания заполняемых элементов.

*Контурные шрифты*. Аналогично векторным, описываются с помощью некоторых математических кривых, но они определяют не символ, а его контур, который заполняется по определенным правилам. Именно этот тип шрифтов и является наиболее популярным.

Для использования векторных и контурных шрифтов необходимо выполнение операции, “создающей” шрифт (заданного рисунка, размера и начертания), годного для отображения. Такая операция называется “растеризацией”. В состав графических оболочек современных операционных систем входят программы - растеризаторы шрифтов определенного формата.

**Наиболее популярные форматы шрифтов** - это TrueTypeFonts (TTF, поддерживается ОС Windows и MacOS) и PostScript (разработан фирмой Adobe, для использования необходима программа Adobe TypeManager). Сейчас на смену этим форматам приходит совместно разработанный этими компаниями формат OpenType.

*Растеризация шрифта* - достаточно ресурсоемкая операция, поэтому контурные шрифты получили распространение только с началом массового применения достаточно мощных компьютеров.

## Структурирование теста

Помимо внешнего вида букв, важное значение имеет пространственное расположение текста. Единицей пространственного размещения служит **абзац**. Как и в литературе, в компьютерном тексте абзацем называется выделенный по смыслу участок.

Для оформления абзаца используют несколько параметров:

***1. Выравнивание*** (выключка) - *правило расположения букв в строке абзаца*. Видов выравнивания четыре: по левому краю, центральное, по правому краю и по ширине полосы набора.

***2. Отступы*** от краев полосы набора.

***3. Абзацный отступ*** (красная строка) - *положение первой строки абзаца*.

***4. Интервалы.***Различают межстрочное расстояние- задается множителем размера шрифта (одинарный, полуторный, двойной интервал) - и промежутки до и после абзаца.

***5. Буквица*** - *крупная выступающая первая буква абзаца*. Часто задается не просто более крупным размером буквы, но и буквой другого рисунка.

Абзацы размещаются в рамках **полосы** - *выделенного участка страницы, как правило, прямоугольной формы, в котором размещаются текст и иллюстрации*. На листе может быть либо одно такое место (одна колонка), либо несколько - тогда говорят о многоколоночном тексте.

Как правило, текстовые процессоры не дают появляться *висячим строкам*-*отдельным строкам абзацев в начале или конце полосы*.

Важным элементом оформления текста на странице являются **поля** - пробелы вдоль края страницы и интервалы между колонками. Для удобного чтения, в силу особенностей восприятия, такие пробелы должны быть обязательно.

Как правило, в достаточно большом (больше нескольких страниц) тексте выделяется несколько смысловых блоков (разделов) и видов содержательного текста - обычный текст, примечания, ссылки и т.п.

Для оформления таких элементов создаются *стили* - *определенные наборы параметров оформления шрифта и абзацев*. Применение стилей позволяет ускорить набор, автоматизировать оформление (например, автоматически создавать оглавления) и изменять внешний вид различных элементов, не разыскивая их по всему тексту. Практически все современные текстовые процессоры опираются на стили, даже если пользователь не использует их. Единство оформления - одно из условий удобочитаемости и красоты печатного издания.

Для решения некоторых типовых задач оформления текстов существующие текстовые процессоры предусматривают два мощных

средства автоматизации (табл.2).

1. **Списки**.

2. **Таблицы**.

Таблица 2

|  |  |
| --- | --- |
| Название | Значение |
| Списки | При оформлении текста это *набор визуально выделенных элементов перечисления*. Элементы выделяют с помощью символа-маркера (**маркированные списки**) либо номером - в упорядоченных списках. При оформлении списка чаще всего также предусматривают форматирование абзацев - так, чтобы они не выступали за маркер. Автоматизация оформления позволяет автоматически маркировать и выделять новые элементы списков. |
| Таблицы | Современные текстовые процессоры предусматривают средства для создания двухмерной структуры размещения информации. Применение таких средств позволяет редактировать структуру и содержание таблицы, добавлять строки и столбцы, изменять их линейные размеры, выделять их с помощью сетки или фона. Фактически каждая ячейка таблицы становится листом в миниатюре. |

Стоит отметить, что файлы текстового процессора содержат массу дополнительных (по отношению к тексту) данных об оформлении и текстовыми очень часто не являются.

Как и текстовые редакторы, текстовые процессоры обладают средствами создания макрокоманд. Современные процессоры реализуют их на развитом языке сценариев, позволяющем решать довольно сложные задачи преобразования и оформления публикаций.

**Примеры программных продуктов:**

Microsoft Word, OpenOffice Writer, StarOffice Word

# 3. Автоматизированная обработка текста

Текстовое представление информации - одно из наиболее удобных для организации автоматической обработки. Связано это с тем, что в этой форме информация представляется в виде близком к исходному языку, что позволяет выполнять преобразования, связанные со смыслом текста.

Существует несколько наиболее распространенных автоматизированных операций, связанных с текстовым представлением.

## 

## Поиск

Задача поиска необходимой информации чаще всего формулируется как поиск фрагментов, содержащих некоторые понятия, в достаточно большом массиве. Большое значение этот вид автоматической обработки получил с ростом популярности межсетевой среды Интернет. Существует несколько подходов к организации такого поиска.

***Первый подход*** опирается на поиск фрагмента текста, соответствующего некоторому образцу. Наиболее популярная форма задания этого образца - так называемые **регулярные выражения**. По сути, это *описание фрагмента текста, удовлетворяющего некоторым условиям*, по тем частям, которые в нем содержатся, и их порядку. Таким способом в большом текстовом массиве можно находить упоминания тех или иных слов, адреса, номера телефонов и т.п. шаблонные элементы.

Достоинство этого подхода - возможность применять его к массиву текста без предварительной обработки. Например, сразу при посимвольном получении текста.

***Второй подход*** предусматривает предварительное создание *специального вида базы для ускорения поиска* - *индекса*. Такой способ применяется для ускорения поиска, если некоторые типовые поисковые запросы повторяются часто или нет возможности формировать/хранить весь массив текста. Например, при организации поисковой машины в среде Интернет.[[4]](#footnote-4)

## Расшифровка или уточнение значений слова

Для решения такой задачи в самых разных видах применяют *словари*- *базы информационных фрагментов, связанных с некоторыми ключевыми словами или словосочетаниями*. Примером таких баз могут быть словари различных языков: англо-русский, русско-английский, толковый и другие виды словарей. Одно из самых распространенных применений словарей - проверка правописания слов при наборе.

Особым видом словарей являются *тезаурусы* - *словари, в которых слова связываются на основе каких-либо лексических отношений*.

Например: слова являются синонимами (смысловыми аналогами), антонимами (противоположны по смыслу) и т.п.

Этот вид словарей важен не только потому, что может помочь при подготовке текстов, но и потому, что это отразит смысл слов - для систем, моделирующих отдельные аспекты мышления человека.

# **Заключение**

Используя закономерности естественного языка и описанные выше средства выполнения некоторых операций и выявления зависимостей, с помощью ЭВМ автоматизируют некоторые операции по смысловому преобразованию текста. Современные системы обработки позволяют создавать краткие обзоры текстов (рефераты) или готовить перевод с одного естественного языка на другой. Приходится отметить, что точного решения эти задачи не имеют, поскольку зачастую трудно подобрать адекватное слово или выражение, учитывая не только формальный перевод, но и грамматические особенности, и культурные. Тем не менее с применением специализированных по областям знания словарей современные системы автоматизированного перевода создают подстрочник, который может дать представление о смысле текста и в дальнейшем помочь переводчику в переводе документа.

**Примеры программных продуктов**

*Системы локального поиска*:

Следопыт, Google Desktop, Microsoft Office Find

*Системы и утилиты автоматизированной обработки текста*:

Grep, lexx, yacc

*Словари*:

Abbyy Lingvo, Multilex

# Список использованной литературы

1. Потапова Р.К. Новые информационные технологии и лингвистика. М., 2002.

2. Хроленко А.Т., Денисов А.В. Современные информационные технологии для гуманитария. Практическое руководство. Москва издательство «Наука», 2007 г.

3. Бовтенко М.А. Компьютерная лингводидактика. Москва: Флинта: Наука, 2008г.

4. Потапова Р.К. Новые информационные технологии и лингвистика. М., 2002.

5. Романенко В. Н., Никитина Г. В. Сетевой информационный поиск: Информация в Интернете; Поисковые машины; Электронные каталоги библиотек; Как формулировать запросы: Практическое пособие. - СПб., 2003

6. Степанов М.Ф. Машинный перевод и общение на естественном языке. Саратов, 2000.

7. Предметный поиск в традиционных и нетрадиционных информационно-поисковых системах: сб. науч. тр. СПб.: Изд-во РНБ, 2001. Вып. 15. 303 c.

8. Денисова Э.В. Информатика. Базовый курс: Учебное пособие. - СПб.: НИУ ИТМО, 2013. - 70 с.

1. Марчук Ю.Н. Основы компьютерной лингвистики. М., 2000. [↑](#footnote-ref-1)
2. Потапова Р.К. Новые информационные технологии и лингвистика. М., 2002. [↑](#footnote-ref-2)
3. Степанов М.Ф. Машинный перевод и общение на естественном языке. – С., 2000. [↑](#footnote-ref-3)
4. Романенко В. Н., Никитина Г. В. Сетевой информационный поиск: Информация в Интернете; Поисковые машины; Электронные каталоги библиотек; Как формулировать запросы: Практическое пособие. - СПб., 2003 [↑](#footnote-ref-4)