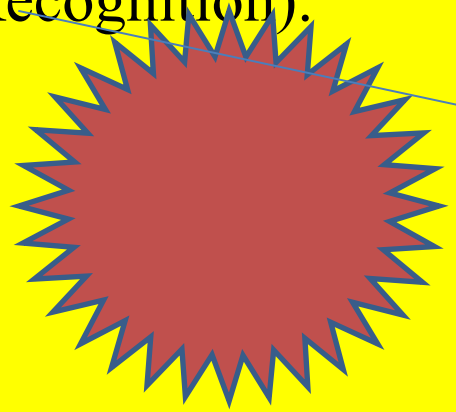
АВТОМАТИЧЕСКОЕ ЧТЕНИЕ ТЕКСТА

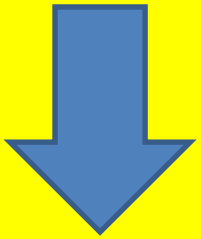
Для быстрого и качественного ввода текстовой информации в компьютер широко используются сканеры.

Общий вид сканера

Для того чтобы «понять» содержание текста, т.е. перевести графическое (точечное) изображение символов в пригодную для дальнейшей обработки (редактирования, реферирования, перевода и т.д.) текстовую форму, необходима система автоматического чтения текста или оптического распознавания символов (OCR-система – Optical Character Recognition).

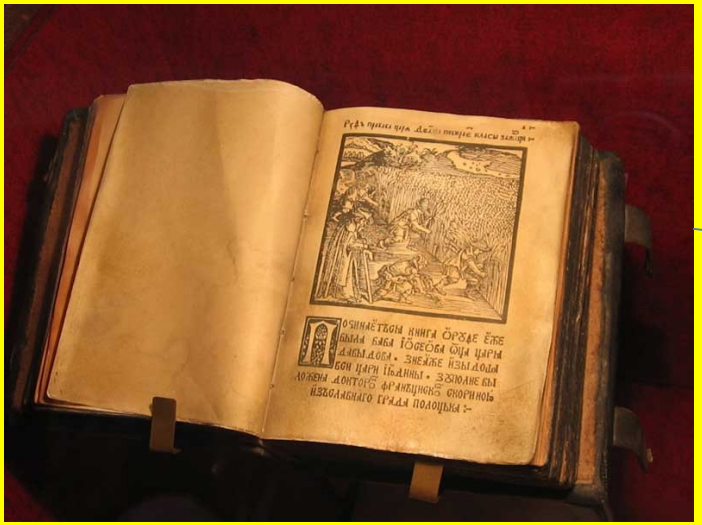


ТЕКСТ



OCR-система

В классическом понимании ***система автоматического чтения текста*** — это компьютерная **программа**, позволяющая **преобразовать** текст с бумажного носителя в электронный текстовый файл, который может быть прочитан средствами обработки текстов.



Автоматическое чтение текста

распознавание речи

решение шахматных задач и головоломок

сочинение музыки и стихотворений

Фрагменты идеи обучения компьютера решать интеллектуальные задачи

К концу 50-х годов эти идеи оформились в отдельную область знания – ***искусственный интеллект.***

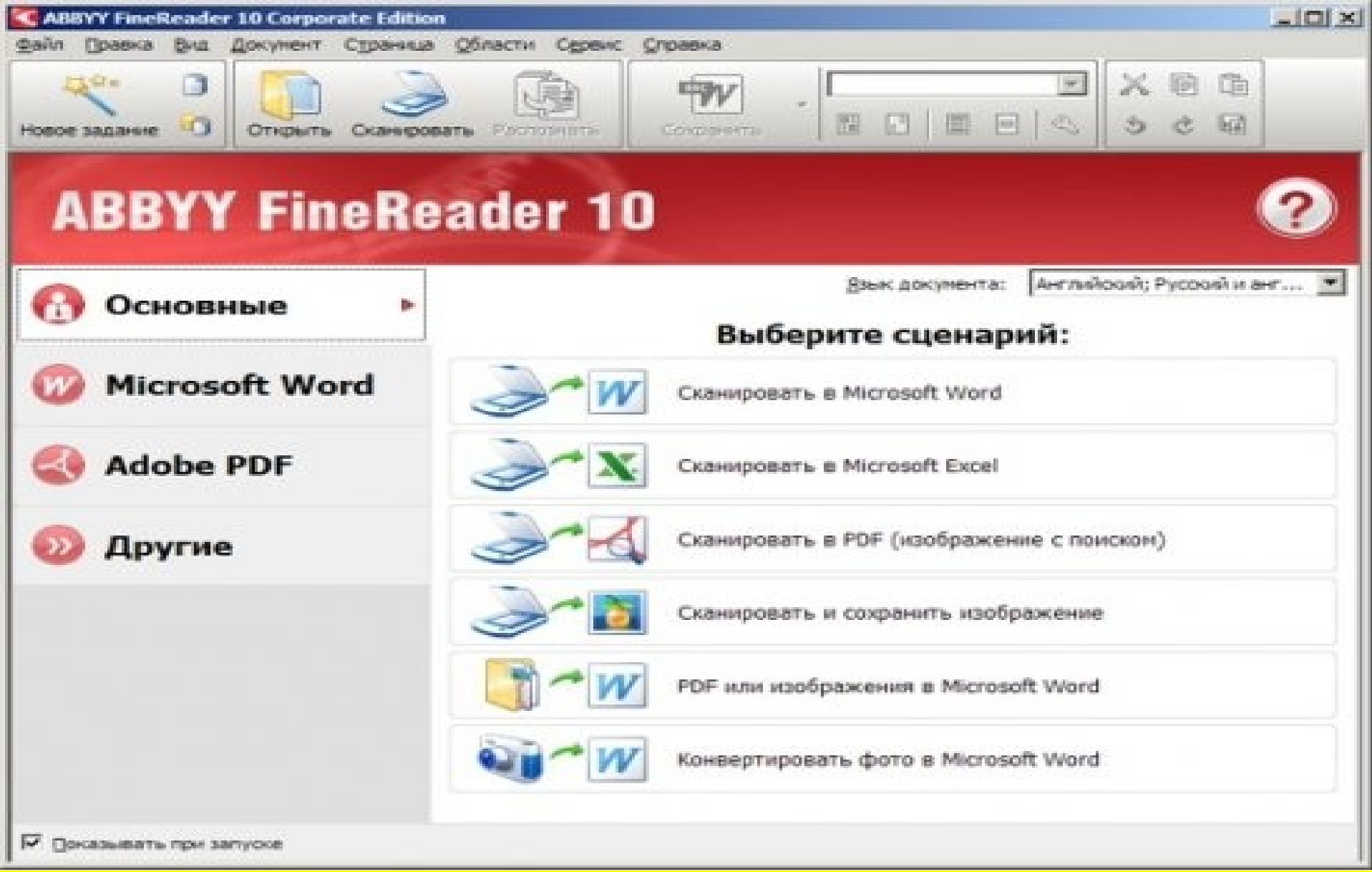
Одной из задач, которая вскоре выделилась в отдельное направление, была задача распознавания образов.

Идеальная компьютерная система распознавания должна уметь

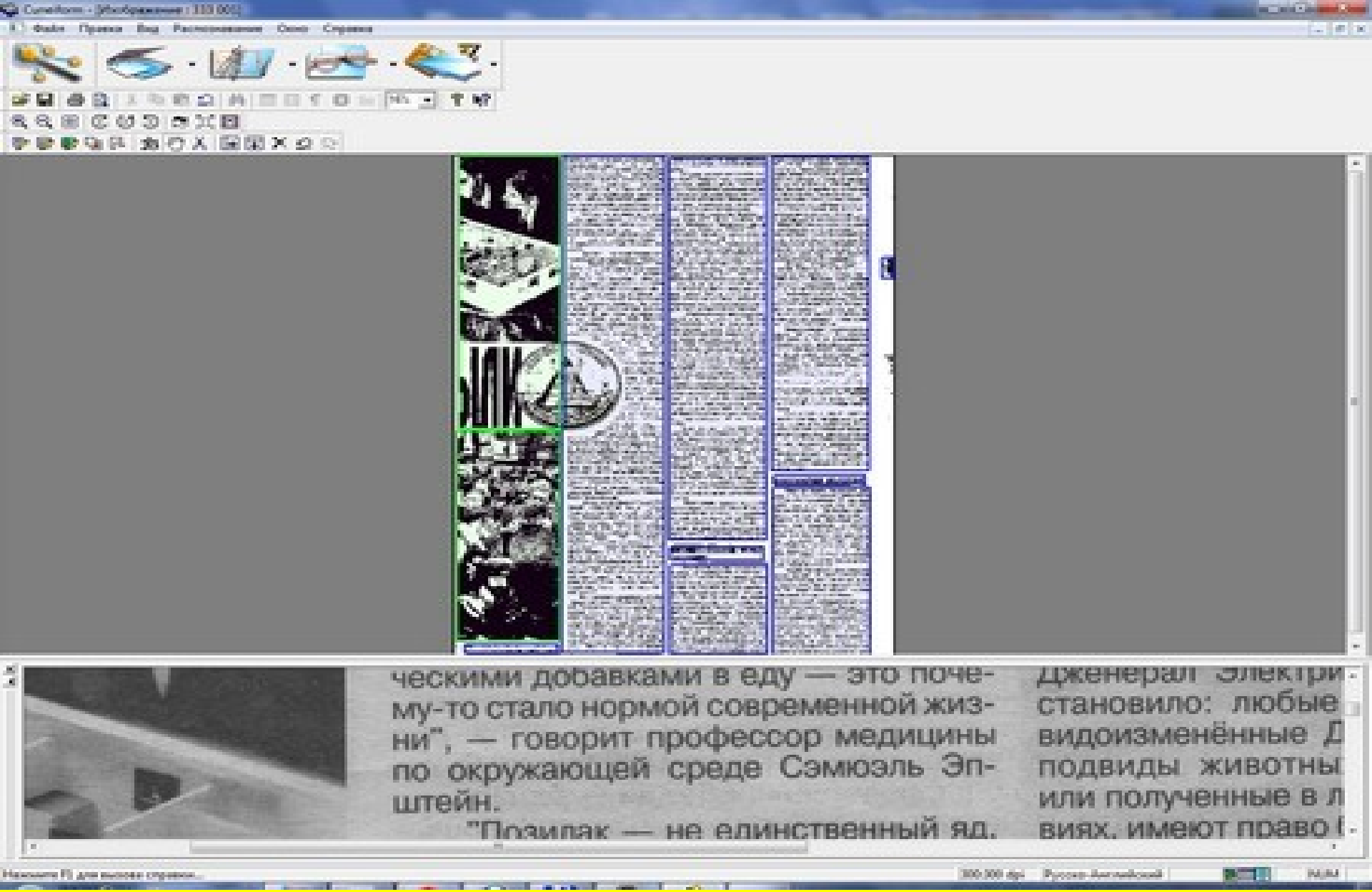
1.формировать,
2.анализировать
3.интерпретировать

любое изображение, в том числе и символьное.

В настоящее время во всем мире широко известны две OCR-системы, созданные российскими разработчиками. Это **FineReader** компании «ABBYY Software House» и **CuneiForm** фирмы «Congitive Technologies».



ABBYY FineReader 10



CuneiForm V12

Основные возможности систем автоматического чтения текста огромны

•В процессе сканирования и распознавания текста документа OCR-системы **автоматически** подбирают **яркость сканирования**, **фрагментируют** каждую страницу, выделяя в ней области **графических иллюстраций и таблиц**, распознают **символы текста**, проверяют **орфографию** распознанных слов и показывают окончательный результат в текстовом редакторе.

•OCR-системы позволяют распознавать печатные символы **почти двух сотен языков**.

•Хорошо распознаются **рукопечатные** символы, т.е. символы, написанные от руки печатными буквами с небольшим интервалом между ними.

•OCR-системы узнают **все используемые в тексте документа шрифты** без предварительного обучения, хорошо воспринимают полужирный, **курсивный**, **слипшийся**, **подчеркнутый** и **многоколоночный** текст. *Изначально в мире преобладали системы автоматического чтения текста, требующие обучения каждому новому шрифту (новой гарнитуре, стилю, размеру и т.д.). Такие системы называли мультифонтовыми (от англ. font — «шрифт»).* Противоположным классом OCR-систем являются так называемые интеллектуальные программы, именуемые еще *омнифонтовыми.* Их не нужно обучать, эти программы распознают разные стилевые начертания одной и той же буквы так как они знают топологию (правила начертания) этой буквы.

•Системы способны **самообучаться** и распознавать **плохо пропечатанные символы** или **символы незнакомых программе языков**.

•Наряду со сплошными текстами (без таблиц и иллюстраций) программы автоматического чтения текста хорошо распознают:

тексты с графикой, подписями, логотипами

таблицы

тексты, напечатанные на цветном (гербовом) фоне

тексты разноформатных документов (например, чертежей)

•OCR-системы поддерживают **все модели сканеров** и **любые графические форматы**.

•Появились и широко используются **сетевые версии** программ автоматического чтения текста.

•Программы автоматического чтения текста поддерживают **публикацию бумажных документов** в глобальной сети Интернет. В процессе распознавания и генерации HTML-страницы ее оформление производится **по всем правилам Web-публикации**.

Точность распознавания OCR-систем на текстах хорошего и среднего качества достигает **97-99 %.**

Развитие программ автоматического чтения текстов в ближайшем будущем пойдет в направлении

повышения точности распознавания текстов низкого качества,

выделения текстовой информации на фоне шумов (например, распознавание номерных знаков автомобилей),

интеграции OCR-систем с различными программами обработки информации (системами машинного перевода, системами автоматического аннотирования и реферирования текстов, электронными архивами, системами автоматизации делопроизводства и т.д.).

Реферат - связный текст, который кратко выражает не только центральную **тему** или **предмет** какого-либо документа, но и **цель**, применяемые **методы**, основные **результаты** описанного исследования или разработки.

Реферат акцентирует внимание читателя на новых сведениях и определяет целесообразность его обращения к исходному документу.

Процесс составления реферата называется **реферированием**.

Аннотацией называют **краткое изложение содержания** документа, дающее общее представление о его теме.

реферат в краткой форме знакомит читателя **с сутью излагаемого в документе содержания** (фактами, методикой, экспериментами и т.п.)	аннотация выполняет лишь сигнальную функцию, сообщая о том, что **опубликована статья или книга на определенную тему**

Процесс составления аннотации называется **аннотированием**.

Рефераты и аннотации представляют собой **вторичные** документы. **Первичные**, или исходные, документы – это книги, статьи, патенты и т.п.

В каждом вторичном документе можно выделить два компонента информации:

Первый компонент содержит информацию **первоисточника** (о чем книга, статья).

Второй компонент – это сведения о самом **первичном** документе (тип документа: книга, статья и т.п.; вид: печатный, рукописный; год издания; место издания и т.д.).

Для оперативного «поверхностного» знакомства с новейшими публикациями используются **рефераты** и **аннотации** книг и статей, которые составляются в специальных организациях и публикуются в реферативных журналах (РЖ) и реферативных сборниках (РС).

Составление реферата или аннотации текста с помощью компьютера называется **автоматическим реферированием или аннотированием**.

При выполнении работы по составлению реферата или аннотации человек (референт) проходит ряд этапов

подготовительный — референт определяет тематическую направленность текста и пытается понять и осмыслить документ в целом

аналитический — референт делит текст на некоторые фрагменты (абзацы, аспекты и т.п.). Каждый фрагмент внимательно изучается, в нем выделяют основные смысловые единицы (предложения, словосочетания, слова). Данный этап заканчивается составлением плана будущих реферата или аннотации

этап непосредственного построения реферата или аннотации – выделенные ранее смысловые единицы (их комбинации или преобразования) располагаются в единый вторичный текст в соответствии с планом реферата или аннотации

В качестве основных смысловых единиц, выделяемых из исходного текста на 2-м этапе, могут выступать:
1)целые ключевые предложения;
2) ключевые словосочетания и слова.

Ключевое (опорное) слово — это термин, относящийся к основному содержанию текста и повторяющийся в нем несколько раз (с учетом всех возможных синонимов).

Ключевое словосочетание — это сочетание слов, среди которых есть одно или несколько ключевых.

Ключевым предложением считается предложение, содержащее два и более ключевых слова или ключевых словосочетания.

Составление плана будущих реферата или аннотации заключается в выделении некоторых смысловых ориентиров, которые на 3-м этапе будут развернуты более подробно.

В качестве подобных ориентиров выступают

основные темы и подтемы исходного текста

основные аспекты исследования

основные ключевые предложения, словосочетания и слова

Создаваемый на 3-м этапе реферат или аннотация содержат выделенные ранее смысловые единицы.

полные (без изменения) ключевые предложения исходного текста

перефразированные ключевые предложения исходного текста

предложения, составленные из ключевых слов или словосочетаний исходного текста с помощью специальных связующих элементов

предложения, обобщающие несколько предложений исходного текста (не обязательно ключевых)

При перефразировании применяются различные лексико-грамматические явления: использование **синонимов**, **конверсивов**, **замен** по принципу «вид — род», «часть — целое» и т.п.

При получении новых предложений из ключевых слов и словосочетаний исходного текста чаще всего используют различные логико-смысловые скрепы, например, *потому что, в то время* ***как, поэтому, вследствие*** **и т. п.**

В обобщающих предложениях исходный текст передается совершенно другими словами. В них то же самое содержание излагается в более кратком виде.

Смысловыми единицами аннотации могут быть:

ключевые слова или словосочетания исходного текста с предшествующими им специальными фразами — реляторами типа: «В статье рассматриваются следующие вопросы:...», «Книга посвящена следующим проблемам: ...» и т.п.

специальные предложения, содержащие оценочные элементы: «Рассматривается важная проблема...», «Статья посвящена актуальной теме...» и т.д.

специальные предложения, содержащие клише, т. е. специализированные словесные штампы, фиксирующие внимание читателя на определенных аспектах содержания: *«Недостаток...* заключается», *«Цель* публикации...», «Ставится *задача...»,* «Делается ***попытка...»*** **и т.д.**

Читая текст повторно (первый раз он читается на подготовительном этапе) или в третий раз, человек мысленно выделяет в нем три типа единиц (предложений, словосочетаний, слов):

единицы, которые **обязательно** должны быть включены в реферат или аннотацию (**новые идеи, гипотезы, новые методы, явления, процессы, новые результаты**), т. е. все новое и оригинальное, что есть в исходном документе. Это и есть основные смысловые единицы текста (ключевые предложения, словосочетания и слова)

единицы, которые отражают фактические данные: параметры изделий, процессов, методов и т.д. Такие единицы не являются принципиально новыми

единицы, которые аргументируют и иллюстрируют единицы первых двух типов

Единицы **первого** уровня обязательно используются при составлении реферата.

Из единиц **второго** уровня используются лишь некоторые (в зависимости от типа реферата или его потребителя).

Третья группа единиц изредка переносится в реферат в обобщенном виде.

Если поручить составление реферата или аннотации компьютеру, то, очевидно, его надо научить выполнять те же действия, которые осуществляет человек.

находить в тексте ключевые слова, словосочетания и предложения

находить в тексте менее значимые единицы

составлять из текстовых единиц двух первых типов смысловые единицы реферата или аннотации

составлять из таких единиц текст реферата или аннотации