

# **Сферы применения систем автоматической обработки текстов**

**Локализация и интернационализация**

Для того чтобы иметь успех на международном рынке, программные продукты должны быть **ЛОКАЛИЗОВАНЫ**

приспособлены к культурным и языковым нормам  
потенциальных покупателей

RU

UK

UA

TR

Во многих программах локализация может быть сравнительно простой

При незначительных изменениях в структуре алгоритма

Меню

Menu

Произошла  
критическая ошибка

A fatal error occured

**При предусмотренной возможности локализации**

Код

Текст

.file

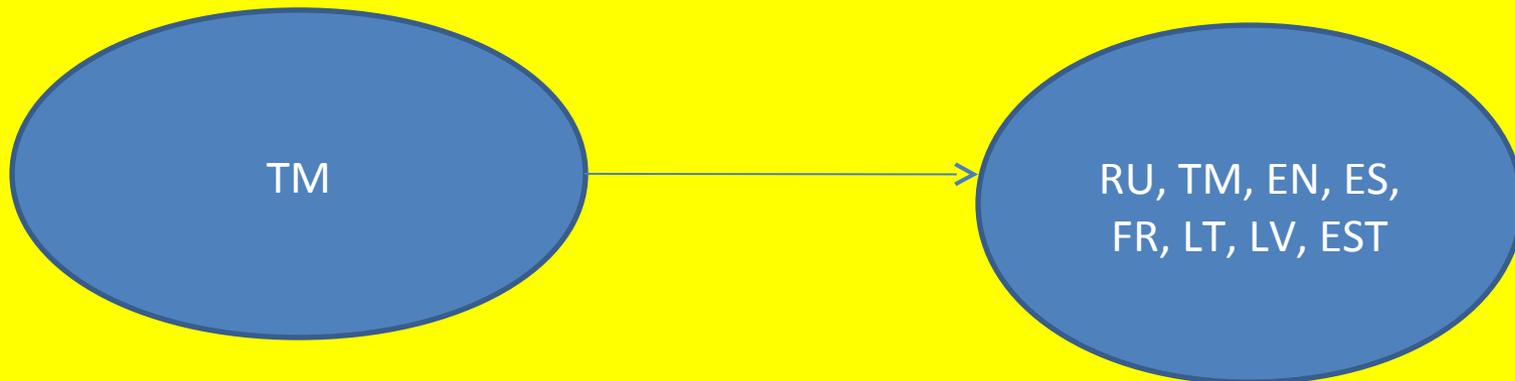
Перевод

Для локализации специализированных лингвистических программ необходимы  
**специализированные словари**  
**более глубокая переработка алгоритма**

**Частично локализацией приходится  
заниматься конечному потребителю**

В идеале программные средства должны быть интернациональными

Купив программу для одного языка, пользователь не должен покупать другую версию для другого языка



Назрела необходимость разработки программ, позволяющих автоматически выбирать язык установки либо предлагать данный выбор.

Данная задача успешно реализуется как в Евросоюзе и США, так и в странах СНГ.

**Работа на ограниченном языке**

Одним из способов разрешения проблем, связанных с обработкой естественного языка, является упрощение и некоторая формализация самих текстов: использование ограниченного языка (подмножества языка).

## Особенности ограниченного (упрощенного) языка

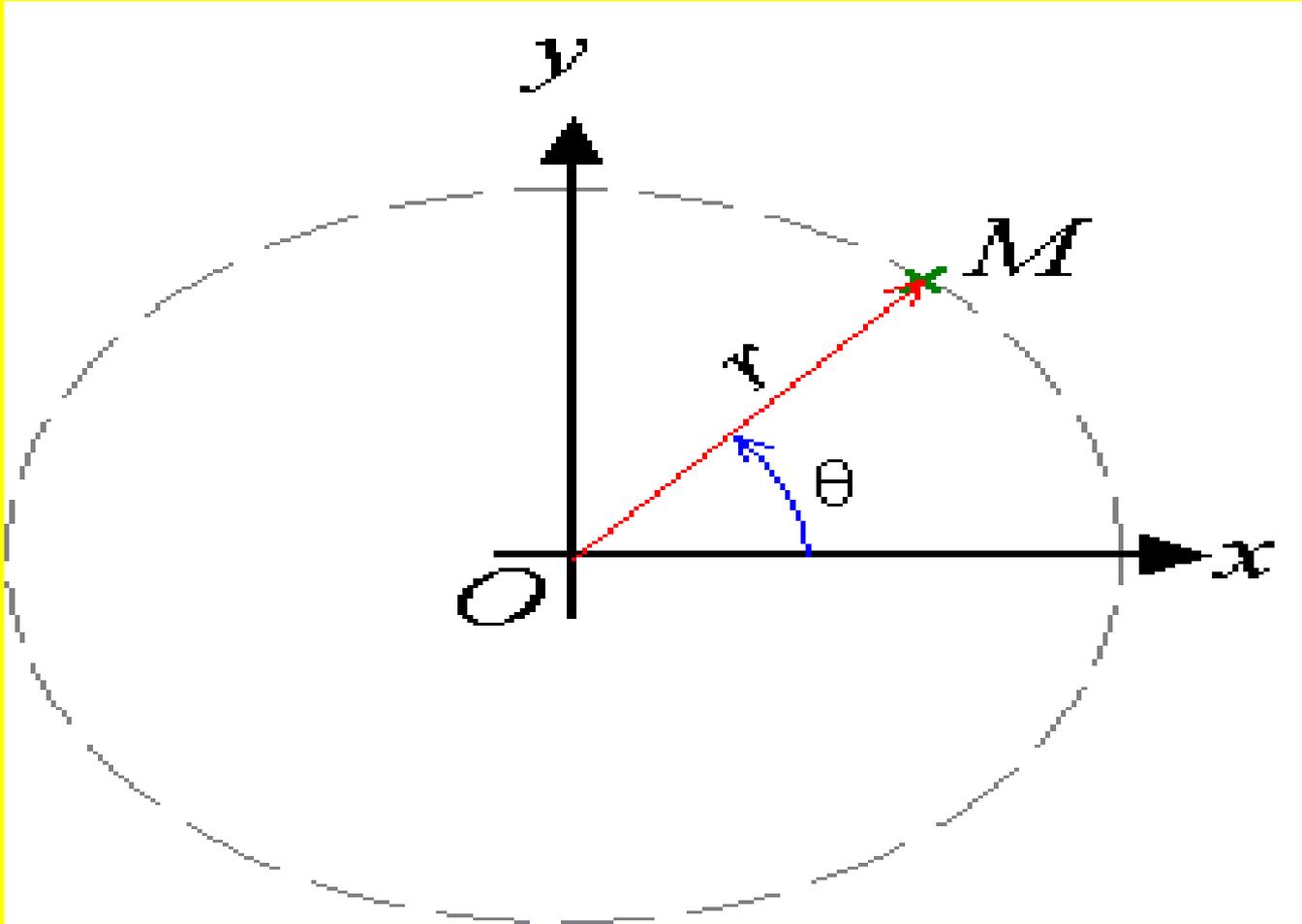
Ограниченный словарь и грамматика, строго определенные предложения

Запрет длинных предложений и цепочек имен существительных (типа "решение проблемы разработки систем перевода на базе представления текста в виде последовательности предложений...")

Запрет пассивных и негативных конструкций, строгие правила в использовании терминов

Тексты должны соответствовать одному из стандартных стилей, быть составленными по определенному шаблону

Достаточно "древним" примером ограниченного языка является BASIC ENGLISH, введенный англичанами для общения с туземным населением в колониях.



Angle



boy



table

<http://ru.wikipedia.org/wiki/Бейсик-инглиш>

Колонизация ввела в быт туземцев множество предметов и понятий, просто не имеющих названий в их родных языках.

Аналогичная ситуация наблюдается во всем мире на сегодняшний день

Например, все специалисты в области компьютерной техники пользуются английскими терминами (*файл, принтер* и т.д.), не пытаясь подыскать эквивалент на родном языке, и мы по-русски говорим *word для windows*, а не *слово для окон*.

## Преимущества ограниченного языка

Документ становится более понятным, удобным для восприятия

Облегчается работа переводчика (меньше возможностей для неоднозначного толкования)

Документ может составить автор, не являющийся носителем языка

Правительства начинают вводить стандарты на подготовку документации, нормы, по которым требуется использование ограниченных языков, особенно в международной торговле.

В связи с этим

Возникает потребность в автоматической проверке на соответствие текста правилам ограниченного языка

Появляется задача создания систем, осуществляющих перевод с естественного языка на ограниченный

Boeing, Caterpillar и несколько других компаний призвали вести всю документацию только на ограниченном языке. Ими разработана система [Boeing Simplified English Checker](#) для проверки соответствия текстов различным промышленным стандартам и государственным нормам. На ее базе создается программа [Clearcheck](#), не только контролирующая правильность текста на ограниченном языке, но и исправляющая ошибки.

Некоторые разработчики прогнозируют создание систем с использованием ограниченных языков, в которых полный и корректный перевод документации будет производиться без вмешательства человека.



Search

Search BR&T

- Boeing Mathematics and Information Software
- Contact Information
- BCSLIB-EXT Software
- BCSLIB Mathematical Software
- 2-stage Nesting Algorithm Software
- Boeing Simplified English Checker**
- VOXMAP Point Shell
- Sparse Optimal Control Software
- IISS

# Boeing Research & Technology

## Simplified English Checker

### Checker for ASD Simplified Technical English Now Available From Boeing

Boeing offers an ASD Simplified Technical English Checker with the following features:

- ▶ **The Boeing Simplified English Checker (BSEC)** helps technical writers check their documents for compliance with [ASD](#) (AeroSpace and Defence Industries Association of Europe) [Simplified Technical English](#), a writing standard for aerospace maintenance documentation.
- ▶ **The BSEC Vocabulary Management System** allows users to add new technical vocabulary and modify advice contained in its internal dictionary and thesaurus.
- ▶ **The BSEC Vocabulary Profiler** gives writers and editors text mining tools such as a word frequency profiler and concordance generator.
- ▶ **The BSEC STE Tutor** helps users improve STE writing skills.
- ▶ **Specification ASD-STE100** in HTML format is included in our product. STE training is available for technical writers.

Contact our Boeing [representatives](#) for more details.

### Quick Links

- ▶ [What is Simplified English?](#)
- ▶ [What is the Simplified English Checker?](#)
- ▶ [Features](#)
- ▶ [Contact Us](#)
- ▶ [Authorized Worldw Resellers](#)

Реклама Boeing Simplified English Checker

**Создание текстовых документов (ввод, редактирование,  
исправление ошибок)**

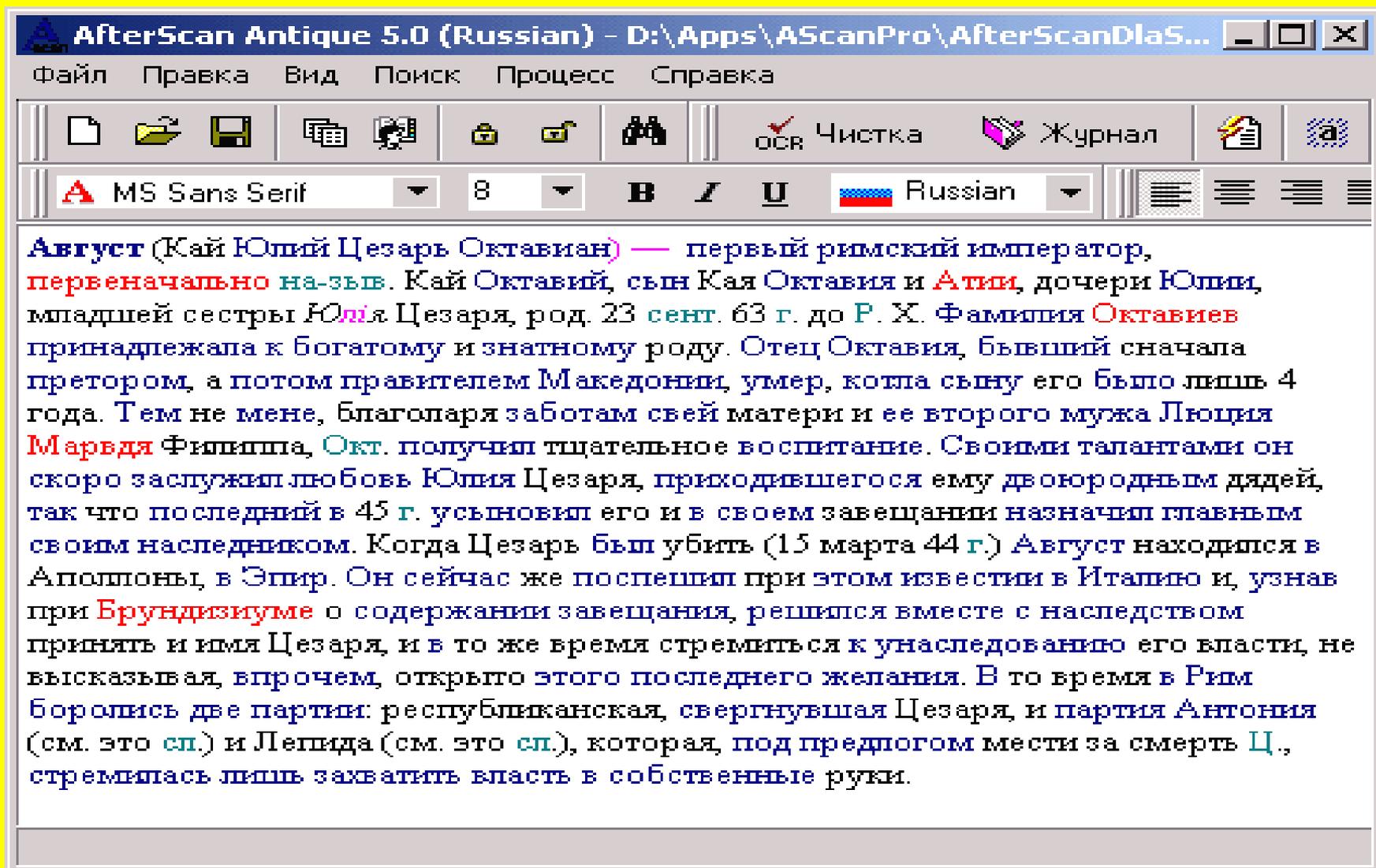
Создание текстовых документов - одна из основных сфер применения персональных компьютеров.

Использование текстовых редакторов обусловлено не только тем, что они облегчают работу, но и тем, что в последнее время во многих сферах деятельности введены стандарты на подготовку текстов, основанные на применении определенных редакторов.

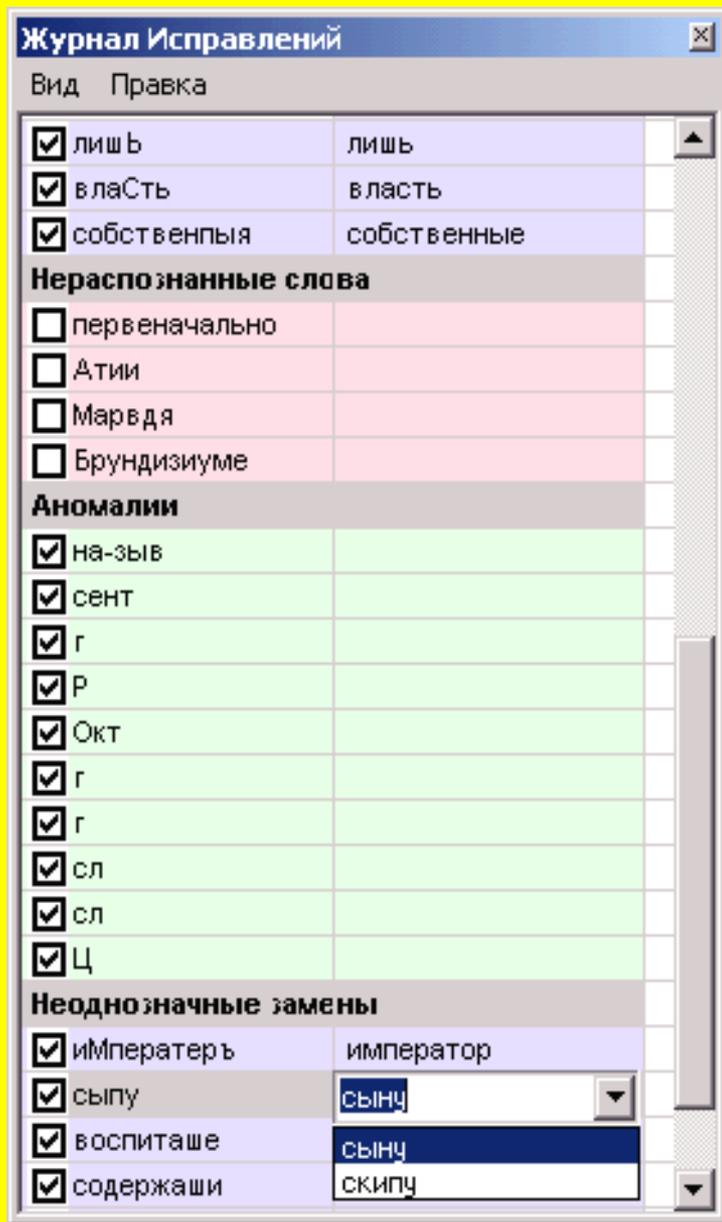
Одним из первых массовых нововведений стало включение в состав текстового редактора программ проверки правописания и внесения необходимых исправлений - ***автокорректоров***.

Чтобы придать своему продукту новые коммерчески перспективные свойства, создатели вынуждены все больше использовать лингвистические знания, применять методы морфологического и синтаксического анализа.

На очереди - создание систем, выполняющих функции научного редактора, т.е. осуществляющих литературную и научную правку текстов, способных производить сложное автоматизированное редактирование текстов на естественном языке.



Автокорректор AfterScan



Журнал исправлений в  
программе  
AfterScan

Проверка текста в таких системах может вестись в режиме **"off-line"** - когда формируется протокол замечаний по тексту, либо в режиме **"on-line"** - когда исправление ошибок ведется по мере их обнаружения (возможно, после получения соответствующего подтверждения от пользователя).

При обнаружении ошибки система может предложить вариант ее исправления (при наличии нескольких вариантов - их упорядоченный список).

Замечания по тексту также могут носить различный характер. Они могут быть **локальными** (указывается фрагмент текста с ошибкой) и **глобальными** (выдается диагностическое сообщение, касающееся всего текста, например: "данный текст труден для восприятия").

# Поиск информации

Часто при поиске информации разного рода (например, аудио- и видео-) работа на самом деле ведется с описаниями на естественном языке (например, для организации поиска фотографий необходимо снабдить каждую из них набором словесных характеристик типа "портрет, профиль, полный рост, женщина", "пейзаж, лес, осень" и т.п.).



Forest.jpg

Очень многие пользователи регулярно сталкиваются с необходимостью быстро просматривать большой объем документов и выбирать из них действительно нужные.

Эта задача возникает при работе с текстовыми базами данных, с электронной почтой, при поиске в Интернете.

Сократить количество просматриваемых документов могут помочь системы ***категоризации***.

При категоризации могут учитываться как чисто внешние показатели документов (объем, расширение имени соответствующего файла и т.п.), так и их содержательные характеристики (название, фамилия автора, ключевые слова), которые могут позволить отнести текст к той или иной тематической рубрике. В последнем случае мы имеем дело с ***рубрицированием*** текстов.

Автоматические методы не настолько совершенны, чтобы создать полноценный реферат путем генерации предложений текста.

Однако уже сейчас возможно **автоматическое реферирование** - составление более или менее информативных и связных рефератов заданного объема (**квазирефератов**) - путем выбора информативных предложений из исходного текста, а также выделение достаточно представительного списка ключевых слов.

base - TextAnalyst

File Edit View Analysis Search Settings Help

Name	Text size	N
SUMMARY	3.23K (16%)	1
Databasing in ...	20.02K	11

Most are compiled by **independent third parties** in the compiled list business or are **by-products** of membership rosters and represent a very lucrative revenue stream for the association.

The second major type of **database** is the publicly available base of inquirers and **customers** rented on the open market by **thousands of businesses**.

Since these **databases** are usually made available by the **owner** of the **product** or the **publication**, **restrictions** are **typically** placed on the use of these names.

And there is usually not much **data** available about these **businesses** and **individuals** because the **owner** does not want to sell anything considered proprietary or sensitive.

For example, see **databases** message took over 100 small **consumer databases** of individuals who had bought

Databasing in the 90's  
[Data](#) and What We're Doing With It!

by Jennifer Barrett  
 Axiom [Corporation](#)

CONWAY, AR –The kind of technological [developments](#) we have seen during the last decade have offered business a virtually unlimited sandbox of [data](#) in which to play. It has stimulated the most creative juices of the business mind, and in many cases resulted in wildly successful [products](#).

The use of [databases](#) in a wide variety of marketing applications and in fulfillment and service to the [customer](#) has become the norm rather than the exception, regardless of how large or small the business. All areas of a business, from [sales](#) and service to accounting and the warehouse, are changing the way they work because of the new [information](#) available to them. Because very little is restricting the collection and use of [data](#) in many [industries](#) today, we see [databases](#) springing up in all shapes and sizes.

Compiled [Databases](#)

The Compiled [Database](#) is one of the simplest forms of a [database](#) today. It appears in two basic flavors, each with very different privacy issues. First is the [database](#) of factual, public [information](#) like census statistics or postal codes, which have few [consumer](#) privacy issues. The use of this [data](#) is usually regulated only by contracts and copyright protection. The second type of compiled [database](#) is one of business or individual and household [information](#). Here we find that the [businesses](#) and the [consumers](#) found on the [database](#) are often not aware of its existence.

Both of these are usually created from some published source, the most common being directories or membership rosters. A few [examples](#) of the kinds of [databases](#) from which you can buy business and individual [information](#) are:

- \* Practicing Attorneys-at-law selectable by type of service and area of law
- \* American Dental Association Members selectable by age, specialty of practice, and geographical location
- \* Corporate and Individual Airplane [Owners](#) and Pilots with type of aircraft and pilot rating .

Done Nodes 91 Documents 1 Total size 20.02K Databasing in the 90 Page 1/11

start Megaputer Intelligenc... base - TextAnalyst v2.3 Samples Document1 - Microsof... 11:16 AM

Окно программы автоматического реферирования TextAnalyst

В качестве **ключевых слов** система может выбирать слова, **наиболее часто встречающиеся в тексте** (и являющиеся при этом **информативными**, т.е. не предлоги, союзы и проч.), либо использовать для отбора какие-либо **синтактико-семантические признаки** (из фрагмента: *"Определение. Интегралом ... называется ..."* можно заключить, что *интеграл* - ключевое слово).

При реферировании из текста отбираются предложения, в наибольшей степени характеризующие его содержание. Таковыми могут считаться, например, **предложения, содержащие ключевые слова** (чем больше, тем лучше), либо **отобранные по некоторым особым признакам**. **Размер реферата (коэффициент сжатия) или количество ключевых слов** задается пользователем.

Результатом работы такой системы может являться некоторый новый текстовый документ (реферат или набор ключевых слов) или же данный документ, в котором ключевые слова или наиболее информативные предложения выделены по тексту.