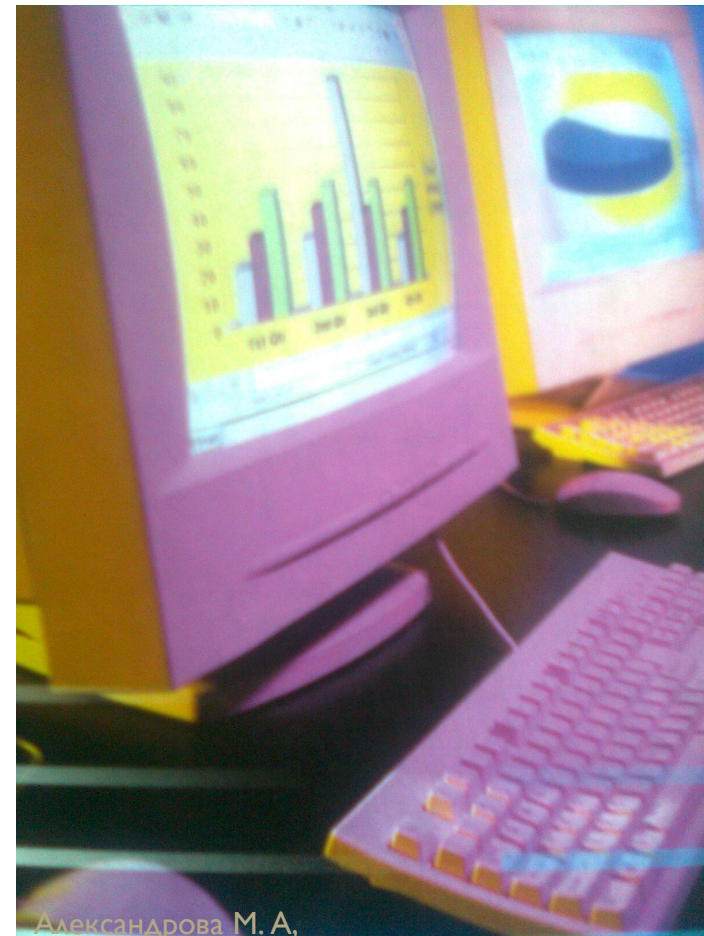


Машинная лексикография

Курс: *«Основы информационных технологий»*

Урок-презентация 4

Составила М. А. Александрова,
студентка 2 курса,
кафедры русской филологии,
БГУ



ОГЛАВЛЕНИЕ

1. Компьютерная лексикография.
2. Направления машинной лексикографии.
3. Машинный словарь.
4. Группы машинных словарей.
 - 4.1. МС для конечного пользователя.
 - 4.2. МС для программ обработки текста.
5. Классификация АСПОТ.
 - 5.1. Классификация по лексическим единицам.
 - 5.2. Классификация по организации словника.
6. Словарь основ.
 - 6.1. Словарь основ: его достоинства и недостатки.
7. Словарь словоформ.
 - 7.1. Словарь словоформ: его достоинства и недостатки.

1. Компьютерная лексикография.

- ▣ *Лексикография* занимается теорией и практикой составления словарей.
- ▣ *Компьютерная лексикография* — совокупность методов и программных средств обработки текстовой информации для создания словарей.

2. Направления машинной лексикографии.

- автоматическое получение из текста с помощью компьютера различных словарей (частотные, конкордансов, терминологические);
- разработка теории и практики составления словарей для МП, ИИ и т. д. (словари корней для морфологического анализа);
- создание машинных версий традиционных словарей.

3. Машинный словарь.

Компьютерный (автоматический) словарь (машинный словарь) – это упорядоченное конечное множество лингвистической информации, представленное в виде некоторой структуры данных, удобной для размещения в памяти ЭВМ и снабженное процедурами автоматического поиска и ведения, т. е. в специальном машинном формате. Такой словарь предназначен для использования на ЭВМ пользователем или компьютерной программой обработки текста.

4. Группы машинных словарей.

Различаются:

- **автоматические словари конечного пользователя-человека (АСКП);**
- **и автоматические словари для программ обработки текста (АСПОТ).**

4.1. МС для конечного пользователя.

Автоматические словари для конечного пользователя, чаще всего являются компьютерными версиями хорошо известных обычных словарей. Они имеют классифицируются подобно традиционным словарям.

4. 2. МС для программ обработки текста .

Их можно назвать автоматическими словарями в точном смысле. Они, как правило, не предназначены для обычного пользователя. Особенности их структуры, сфера охвата словарного материала задаются теми программами, которые с ними взаимодействуют. Существует несколько классификаций МС этой группы по типам.

5. Классификация АСПОТ.

- Классификация по характеру *лексических единиц*, входящих в словарь;
- классификация по способу *организации словника*.

5. 1. Классификация по лексическим единицам.

- синтаксический словарь;
- семантический словарь;
- словарь морфем;
- псевдооснов;
- словарь устойчивых словосочетаний;
- словари основ (список основ и окончаний);
- словари словоформ и т.д.

5.2. Классификация по организации словника.

- частотные
- алфавитные;
- тезаурусы (используемые в системах информационного поиска);
- конкордансы (группировка по ключевым словам);
- переводные словари.

6. Словарь основ.

Словарь основ - машинный словарь, состоящий из списка основ и списка окончаний. Каждой основе и каждому окончанию дается соответствующий код морфологического класса и код лексико-грамматической информации. Во время работы такого словаря необходимые формы слов образуются путем присоединения по заданным правилам соответствующих окончаний к основам.

6.1. Словарь основ: его достоинства и недостатки.

Достоинства: сокращение объема занимаемой МС памяти.

Недостатки: усложнение морфологического анализа и описания ЕЯ.

(Нередко экономия памяти в словарях основ является неоправданной за счет громоздких и не всегда эффективных алгоритмов анализа).

7. Словарь словоформ.

Словарь словоформ состоит из всех словарных форм определенного ЕЯ. Для русского и белорусского языков, являющихся флективно-богатыми языками, при построении МС словоформ более приемлемой является так называемая *гнездовая структура*.

Под гнездом подразумевается совокупность словоформ одной основы (множество всех грамматических форм некоторого основного слова). При использовании такой структуры, в памяти ЭВМ явно хранится лишь основное слово, а для остальных - информация о том насколько они отличаются от него.

7.1. Словарь словоформ: его достоинства и недостатки.

Достоинства: значительно упрощается морфологический анализ;

Недостатки: требуют больше памяти для размещения.

(Но: ресурсы современных ЭВМ позволяют хранить словари практически любых необходимых размеров, поэтому использование словарей словоформ предпочтительнее).