МИНОБРНАУКИ РОССИИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

“ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ”

Факультет романо-германской филологии

Кафедра немецкой филологии

Курсовая работа

Компьютерная лексикография

Студент Малашенкова Е.А. 27.09.2018

Руководитель Донина О.В. к.т.п.л., кандидат филологических наук 27.09.2018

2018

## Содержание

* + [Введение](https://ru.wikiversity.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%BB%D0%B5%D0%BA%D1%81%D0%B8%D0%BA%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D1%8F#%D0%92%D0%B2%D0%B5%D0%B4%D0%B5%D0%BD%D0%B8%D0%B5)
* [1. Лексикография](https://ru.wikiversity.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%BB%D0%B5%D0%BA%D1%81%D0%B8%D0%BA%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D1%8F#%D0%9B%D0%B5%D0%BA%D1%81%D0%B8%D0%BA%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D1%8F)
* 1.1. Теоретическая лексикография
* 1.2. Практическая лексикография
* 2. Компьютерная лексикография
* 2.1.[Особенности компьютерной лексикографии](https://ru.wikiversity.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%BB%D0%B5%D0%BA%D1%81%D0%B8%D0%BA%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D1%8F#%D0%9E%D1%81%D0%BE%D0%B1%D0%B5%D0%BD%D0%BD%D0%BE%D1%81%D1%82%D0%B8_%D0%BA%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BD%D0%BE%D0%B9_%D0%BB%D0%B5%D0%BA%D1%81%D0%B8%D0%BA%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D0%B8)
* 2.2. Инструментальные средства компьютерной лексикографии
* 2.3. История компьютерной лексикографии
* 2.4. Основные понятия компьютерной лексикографии
* 2.5. Электронные словари
* 2.6. Онлайн- словари
* 2.7. Извлечение информации
* 2.8. Распознавание и синтез речи
* 2.9. Задачи компьютерной лексикографии
* Заключение
* Список литературы

**Введение**

В нашей повседневной жизни мы всё чаще и чаще сталкиваемся с термином “электронный словарь”, который сразу дает нам понять, что поиск слов будет проходить не с использованием бумажного носителя. Обычно подразумевается, что словарь на компьютере - это введенный в него бумажный словарь, снабженный удобными средствами поиска и отображения.

Компьютерная лексикография как область прикладной лингвистики, производящая такие словари, оказывается лишенной собственного языкового предмета. На ее долю оставляется только эффектная демонстрация канонического содержания.

Я бы хотела показать, что компьютерная лексикография - особая отрасль лексикографии со своими собственными подходами и решениями по содержанию и созданию словаря.

Электронный словарь - это особый лексикографический объект, в котором могут быть реализованы и введены в обращение многие продуктивные идеи, не востребованные по разным причинам в бумажных словарях.

Необходимо сразу оговориться, что речь идет о тенденциях, потенциальных возможностях компьютерной лексикографии, часть которых еще не реализована или даже еще не осознана.

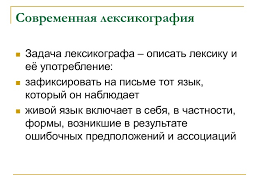
Прежде чем говорить о “компьютерной лексикографии” , нужно немного разобраться, что же такое “лексикография”.

**1.Лексикография**

[Лексикография](https://ru.wikipedia.org/wiki/%D0%9B%D0%B5%D0%BA%D1%81%D0%B8%D0%BA%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D1%8F) (от гр. lexikon — словарь, grapho — пишу) изучает теорию и практику составления различных языковых словарей.

Так же ещё лексикографию относят к разделу языкознания, занимающимся практикой и теорией составления *словарей*.

Иначе говоря, лексикографией называется, во-первых, "наука о составлении словарей" и, во-вторых, "составление словарей как описание лексики данного языка". Кроме того, этим термином обозначается "совокупность словарей, относящихся к данному языку или области знания".



1. Задачи современного лексикографа

**1.1. Теоретическая лексикография**

**Теоретическая лексикография** изучает вопросы разработки макроструктуры и микроструктуры словаря, разработку типологии словарей, историю развития лексикографии.

Макроструктура словаря определяет отбор лексики, объём и характер словника, принципы расположения материала.

Микроструктуры словаря определяет структуру словарной статьи, типы словарных определений, соотношение разных видов информации о слове, типы языковых иллюстраций.

**1.2. Практическая лексикография**

**Практическую лексикографию** можно определить как процесс составлением словарей различных типов на базе теоретических разработок.

|  |
| --- |
| Этапы разработки словаря: |
| 1. Разработка системы требований, касающихся назначения и круга пользователей. |
| 1. Разработка системы требований, касающихся таких параметров словаря как единицы описания, объем, структуру, вид словарной информации. |
| 1. Разработка текстов, расписывание контекстов, характеристика грамматических форм, составление предварительных словников. |
| 1. Дистрибутивный анализ текстов, тесты с носителями языка. |
| 1. Обобщение экспериментальных данных. |
| 1. Построение дефиниций на соответствующем метаязыке и их проверка в ходе новых экспериментов. |
| 1. Сбор и систематизация дополнительной информации о каждой языковой единице. |
| 1. Оформление словарных статей. |
| 1. Системный анализ и упорядочение словарных статей |
| 1. Оформление словаря |

**2.Компьютерная лексикография**

**Компью́терная лексикогра́фия** — прикладная [научная дисциплина](https://ru.wikipedia.org/wiki/%D0%9D%D0%B0%D1%83%D1%87%D0%BD%D0%B0%D1%8F_%D0%B4%D0%B8%D1%81%D1%86%D0%B8%D0%BF%D0%BB%D0%B8%D0%BD%D0%B0) в [языкознании](https://ru.wikipedia.org/wiki/%D0%AF%D0%B7%D1%8B%D0%BA%D0%BE%D0%B7%D0%BD%D0%B0%D0%BD%D0%B8%D0%B5), которая изучает методы использования компьютерной техники для составления словарей. Это временная дисциплина периода перехода от ручной и рукописной [лексикографической](https://ru.wikipedia.org/wiki/%D0%9B%D0%B5%D0%BA%D1%81%D0%B8%D0%BA%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D1%8F) практики к новым безбумажным [информационным технологиям](https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D1%86%D0%B8%D0%BE%D0%BD%D0%BD%D0%B0%D1%8F_%D1%82%D0%B5%D1%85%D0%BD%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F).

Компьютерная лексикография представлена совокупностью методов и программных средств обработки текстовой информации для создания [словарей](https://ru.wikipedia.org/wiki/%D0%A1%D0%BB%D0%BE%D0%B2%D0%B0%D1%80%D1%8C). В рамках компьютерной лексикографии разрабатываются компьютерные технологии составления и эксплуатации словарей. Специальные программы — [базы данных](https://ru.wikipedia.org/wiki/%D0%91%D0%B0%D0%B7%D0%B0_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85), компьютерные картотеки, программы обработки текста — позволяют в автоматическом режиме формировать словарные статьи, хранить словарную информацию и обрабатывать её.

Множество различных компьютерных лексикографических программ разделяются на две большие группы: программы поддержки лексикографических работ и электронные словари различных типов, включающие лексикографические базы данных.

**2.1.Особенности компьютерной лексикографии**

В настоящее время лексикография находится под сильным воздействием новых методов обработки информации.

Изменение инструментальных средств приводит к появлению новых словарных технологий.

Современная информационная технология лексикографии — компьютерная лексикография. Значительная часть интеллектуальных операций переходит в разряд рутинных. При этом наблюдается процесс перехода части лексикографов осваивает новые профессии и отходит от "чистой" лексикографической деятельности, и начинают заниматься издательской деятельностью или организаторами лексикографических исследований и издателями их результатов, с другой стороны, часть специалистов, прежде всего, информатики активно занимаются лексикографической деятельностью.

Компьютерная лексикография представляет собой:

* быстро развивающуюся отрасль компьютерной индустрии;
* прикладную научную дисциплину в языкознании, изучающую методы, технологию и отдельные приемы использования компьютерной техники в теории и практике составления словарей.

Компьютерная лексикография — совокупность методов и программных средств обработки текстовой информации для создания словарей.

#### **2.2.Инструментальные средства компьютерной лексикографии**

К инструментальным средствам в рамках компьютерной лексикографии относятся – базы данных, компьютерные картотеки, программы обработки текста. Эти программы позволяют в автоматическом режиме формировать словарные статьи, хранить словарную информацию и обрабатывать ее.

Множество различных компьютерных лексикографических программ разделяются на две больших группы:

* программы поддержки лексикографических работ,
* компьютерные (автоматические) словари различных типов, включающие лексикографические базы данных.

Компьютерный (автоматический) словарь – это словарь в специальном машинном формате, предназначенный для использования на ЭВМ пользователем или компьютерной программой обработки текста.

Таким образом различаются автоматические словари конечного пользователя-человека и автоматические словари для программ обработки текста.

Автоматические словари, предназначенные для конечного пользователя, по интерфейсу и структуре словарной статьи существенно отличаются от автоматических словарей, включенных в системы машинного перевода, системы автоматического реферирования, информационного поиска и т.д.

Это могут быть компьютерные версии известных обычных словарей. Например: компьютерные аналоги толковых словарей английского языка (автоматический Вебстер, автоматический толковый словарь английского языка издательства Коллинз, автоматический вариант Нового большого англо-русского словаря под ред. Ю.Д.Апресяна и Э.М.Медниковой), компьютерная версия словаря Ожегова.



Автоматические словари для программ обработки текста 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.

## **2.3.История компьютерной лексикографии**

Термин **«Компьютерная лексикография»** был придуман для обозначения области изучения электронных словарей и появился во второй половине 20 века. Новаторские работы Эвенса и Амслера послужили толчком для дальнейшего изучения электронных словарей. Они провели исследование ,используя Седьмой Энциклопедический словарь Вебстера. Толчком для этих ислледований послужило широкое использование Longman Dictionary of Contemporary English, который и в наши дни остаётся одним из самых лучших электронных словарей.

Первоначально электронные словари имели такую же форму записи, как обычные словари, и исследователям приходилось тратить много времени для интерпретации такой формы записи (например, чтобы определить, к какой части речи относится определенное слово). С развитием технологий издатели решили отделить базу данных электронного словаря от того, как он выглядит при печати. Сегодня существуют более удобные формы записи, например расширяемый [язык разметки XML](https://ru.wikipedia.org/wiki/XML). Используя XML, исследователи получают быстрый доступ к информации, хранящейся в электронном словаре.

## **2.4.Основные понятия компьютерной лексикографии**

* ***Автоматический словарь*** — это словарь в специальном машинном формате, предназначенный для использования на ЭВМ пользователем или компьютерной программой обработки текста. Иными словами, различают [автоматические словари пользователя-человека](https://ru.wikipedia.org/wiki/%D0%AD%D0%BB%D0%B5%D0%BA%D1%82%D1%80%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D1%81%D0%BB%D0%BE%D0%B2%D0%B0%D1%80%D1%8C) и [автоматические словари для программ обработки текста](https://en.wikipedia.org/wiki/Machine-readable_dictionary). Автоматические словари, предназначенные для человека, по интерфейсу и структуре словарной статьи существенно отличаются от автоматических словарей, включённых в системы машинного перевода, системы автоматического реферирования, информационного поиска и т. д.
* [***Гипертекст***](https://ru.wikipedia.org/wiki/%D0%93%D0%B8%D0%BF%D0%B5%D1%80%D1%82%D0%B5%D0%BA%D1%81%D1%82) — это множество текстов со связывающими их отношениями.

Гипертекстовые технологии позволяют легко сочетать различные виды информации — обычный текст, рисунок, график, таблицу, схему, звук и движущееся изображение. Как традиционный текст, так и гипертекст — феномены, порожденные новыми технологиями. В первом случае технология позволила легко тиражировать и распространять знания самых различных типов, а во втором — компьютерные технологии дали возможность изменить сам внешний вид текста и его структуру. Разнородность гипертекста — это первое технологическое свойство гипертекста, технологическое в том смысле, что оно непосредственно следует из используемой компьютерной технологии. Второе технологическое свойство гипертекста — его нелинейность. Гипертекст не имеет стандартной, обычной последовательности чтения. Прочие свойства гипертекста в той или иной степени являются следствиями из этих двух технологических свойств.

Практическая ценность гипертекста в компьютерной лексикографии состоит в том, что он описывает тип интерактивной среды с возможностями переходов по ссылкам. Нелинейный характер гипертекста дает возможность формирования информации в виде разветвлённой структуры, что позволяет в значительной степени расширить рамки словарной статьи, то есть он описывает тип интерактивной среды с возможностями перехода по ссылкам. Слова, фразы или рисунки, которые являются ссылками на тот или иной текст или рисунок, позволяют пользователю выбрать необходимую информацию и просмотреть связанные с ней сведения и материалы.

* ***Средства навигации по словарю*** — ссылки, внедрённые в различные элементы электронной среды — часть гипертекстового устройства электронного словаря, представляющего собой соединение смысловой структуры, структуры внутренних связей некоего содержания и технической среды и технических средств, дающих человеку возможность осваивать структуру смысловых связей, а также осуществлять переходы между взаимосвязанными элементами[[12]](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%BB%D0%B5%D0%BA%D1%81%D0%B8%D0%BA%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D1%8F#cite_note-%D0%A2%D0%B0%D0%B1%D0%B0%D0%BD%D0%B0%D0%BA%D0%BE%D0%B2%D0%B0_%D0%92.%D0%94-12).

## **2.5.Электронные словари**

***Электронный словарь*** — это любой упорядоченный, относительно конечный массив лингвистической информации, представленный в виде списка, таблицы или перечня, удобного для размещения в памяти ЭВМ и снабженного программами автоматической обработки и пополнения..

Термин **электронный словарь** может быть использован для обозначения любого справочного материала, хранящегося в электронном виде и предоставляющего информацию о написании, значении или использовании слов. Таким образом, [система проверки правописания](https://ru.wikipedia.org/wiki/%D0%A1%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B0_%D0%BF%D1%80%D0%BE%D0%B2%D0%B5%D1%80%D0%BA%D0%B8_%D0%BF%D1%80%D0%B0%D0%B2%D0%BE%D0%BF%D0%B8%D1%81%D0%B0%D0%BD%D0%B8%D1%8F) в текстовом редакторе, устройство, которое сканирует и переводит напечатанные слова и электронная версия бумажного словаря — всё это электронные словари, имеющие сходные системы хранения и поиска.

Наиболее известные электронные словари:

* Collins Cobuild Student’s Dictionary
* [Cambridge International Dictionaries](http://dictionary.cambridge.org/)
* [Оксфордский словарь английского языка](https://ru.wikipedia.org/wiki/%D0%9E%D0%BA%D1%81%D1%84%D0%BE%D1%80%D0%B4%D1%81%D0%BA%D0%B8%D0%B9_%D1%81%D0%BB%D0%BE%D0%B2%D0%B0%D1%80%D1%8C_%D0%B0%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%BE%D0%B3%D0%BE_%D1%8F%D0%B7%D1%8B%D0%BA%D0%B0)

### **2.6.Онлайн-словари**

Существует большое количество веб-сайтов, работающих в качестве онлайн-словарей, обычно специализирующихся в каком-либо направлении. Некоторые из них содержат только те данные (часто включая неологизмы), которые были добавлены самими пользователями. Вот несколько наиболее известных примеров:

* [Dictionary.com](https://en.wikipedia.org/wiki/Dictionary.com)
* [Double-Tongued Dictionary](https://en.wikipedia.org/wiki/Double-Tongued_Dictionary) (данные добавлены пользователями)
* [Free On-line Dictionary of Computing](https://ru.wikipedia.org/wiki/Free_On-line_Dictionary_of_Computing)
* [LEO (website)](https://en.wikipedia.org/wiki/LEO_(website))
* [Logos Dictionary](https://en.wikipedia.org/wiki/Logos_Dictionary)
* [Pseudodictionary](https://en.wikipedia.org/wiki/Pseudodictionary) (только юмористические неологизмы, добавленные пользователями)
* [Urban Dictionary](https://ru.wikipedia.org/wiki/Urban_Dictionary) (словарь англоязычного сленга)
* [WWWJDIC](https://en.wikipedia.org/wiki/WWWJDIC) (японский онлайн-словарь)
* [Визуальный словарь](http://www.vslovar.ru/) (для каждого слова строится его понятийное окружение).
* [Викисловарь](https://ru.wikipedia.org/wiki/%D0%92%D0%B8%D0%BA%D0%B8%D1%81%D0%BB%D0%BE%D0%B2%D0%B0%D1%80%D1%8C)

.

### **2.7.Извлечение информации**

[Извлечение информации](https://ru.wikipedia.org/wiki/%D0%98%D0%B7%D0%B2%D0%BB%D0%B5%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B8%D0%BD%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D1%86%D0%B8%D0%B8) - это задача автоматического извлечения структурированных данных из неструктурированных или слабо структурированных или слабо структурированных машиночитаемых документов. Проблеме извлечения информации была обозначена на[Message Understanding Conferences](https://en.wikipedia.org/wiki/Message_Understanding_Conferences), где основной задачей было извлечь из текста определённые данные и поместить в заданные слоты шаблонов. Заполнение шаблонов не требует полного разбора текста, этого можно достигнуть путём сопоставления с неким образцом(например, с помощью регулярных выражений). Слоты шаблонов заполняются серией слов, обычно классифицированных. Например, имена людей, названия организаций, химические элементы и т. д.

Для извлечения имён людей, например, могут применяться шаблоны, использующие электронные словари, содержащие списки имён и сокращений, предшествующих именам людей. Часто списки могут быть очень большими, например такие, как список названий компаний или записи географического справочника. Названия можно определить достаточно надежно, не выходя за рамки простых списков, так как в тексте они появляются в качестве просто существительных. Распознать и охарактеризовать событие в тексте тоже можно с помощью такой модели, но необходимо использование дополнительный лексической информации.

События обычно описываются глаголами, и это описание может быть выражено в виде различных синтаксических шаблонов. Несмотря на то, что эти модели могут быть выражены с некоторой степенью достоверности (например, компания наняла человека или человек был нанят компанией) в качестве основы для сравнения строк, этот подход не позволяет достичь желаемого уровня общности. Распознание события влечет за собой частичный разбор предложения.

Большей общности можно достигнуть путём расширения шаблонов требуемых семантических классов. Электронный словарь [WordNet](https://ru.wikipedia.org/wiki/WordNet) широко используется в IE, в частности, с использованием [гиперонимических](https://ru.wikipedia.org/wiki/%D0%93%D0%B8%D0%BF%D0%B5%D1%80%D0%BE%D0%BD%D0%B8%D0%BC) отношений как основы для определения семантических классов. Дальнейшее развитие в IE, вероятно, будет сопровождаться использованием более сложных вычислительных словарей.

### **2.8.**[**Распознавание**](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D0%BE%D0%B7%D0%BD%D0%B0%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D1%80%D0%B5%D1%87%D0%B8) **и** [**синтез**](https://ru.wikipedia.org/wiki/%D0%A1%D0%B8%D0%BD%D1%82%D0%B5%D0%B7_%D1%80%D0%B5%D1%87%D0%B8) **речи**

Использование электронных словарей в технологиях [распознавания речи](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D0%BE%D0%B7%D0%BD%D0%B0%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D1%80%D0%B5%D1%87%D0%B8) ограничено. Машиночитаемые словари обычно содержат произношение, но эта информация обеспечивает только первый шаг в решении проблемы распознавания и синтеза речи. Электронный словарь речевой лексики включает в себя орфографическую форму слов или каноническое произношение. Словарь полных форм также содержит в словарной статье все формы слов; форм могут генерироваться на основе правил, но обычно все формы слов просто хранятся в словаре.

Знания канонического произношения недостаточно для обработки разговорного языка. Необходимо учитывать варианты произношения, являющиеся результатом областных различий, влияние родного языка для иностранцев, зависимость произношения и ударения от порядка слов. Некоторые из этих трудностей можно решить алгоритмически, но решение большинства из них возможно только благодаря более обширному набору информации. В результате речевые базы данных содержат эмпирические данные о фактическом произношении, фрагменты разговорной речи и её нотацию в письменной форме. Эти базы данных включают информацию о тех, чьи голоса записаны, типе речи, качестве записи и прочие данные. Самое главное, что эти базы данных содержат речевые данные в виде сигнала, записанного в аналоговой или цифровой форме. В связи с большими объёмами данных, участвующих в реализации основных систем распознавания и синтеза речи, эти системы ещё не содержат полный спектр семантических и синтаксических возможностей обработки озвученных данных.

## Преимущества электронных словарей над бумажными :

|  |  |
| --- | --- |
| **Многофункциональность** | разнообразие дополнительных функций, которые упрощают обращение к словарю. Например, можно указать части речи, происхождение, а так же словообразование, табуированность лексики |
| **Использование средств мультимедиа** | озвучивание заголовочных слов, введение иллюстративного материала с фотографиями, анимацией, видеофрагментами |
| **Актуальность и динамичность** | возможность постоянного обновления информации, а так же изъятия устаревших данных |
| **Большой объем словарной базы** | У большинства электронных словарей терминологическая база превышает базу бумажных словарей |
| **Вариативность в использовании** | возможность использования словарей в локальной и глобальной сетях |
| **Универсальность** | как правило, программы позволяют работать сразу с несколькими языками и направлениями перевода |
| **Удобный поиск** | возможность использования эффективной системы поиска ([полнотекстовый поиск](https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%BB%D0%BD%D0%BE%D1%82%D0%B5%D0%BA%D1%81%D1%82%D0%BE%D0%B2%D1%8B%D0%B9_%D0%BF%D0%BE%D0%B8%D1%81%D0%BA), одновременный поиск в нескольких словарях, высокая скорость поиска) |

## **2.9.Задачи компьютерной лексикографии**

## Перед учёными и программистами при преобразовании бумажных словарей в машиночитаемые словари встает множество проблем:

|  |  |
| --- | --- |
| [**Парсинг**](https://ru.wikipedia.org/wiki/%D0%9F%D0%B0%D1%80%D1%81%D0%B8%D0%BD%D0%B3) | задача, которая заключается в извлечении информации из словаря, написанного на естественном языке и представление её в виде древовидной структуры с узлами, которые состоят из атрибутов. Каждый атрибут имеет своё значение |
| **Создание лексической базы данных** | задача состоит в том, что при создании лексической базы данных, которая хранит информацию формируемого словаря, необходимо обеспечить быстрый, гибкий и удобный доступ. Так же записи, которые будут здесь храниться, должны иметь структурированный вид. Язык запросов к данной базе должен обеспечивать наиболее удобное добавление, изменение и удаление информации |
| **Лексический анализ электронных словарей** | задача состоит в том, чтобы выполнить анализ содержимого словаря (лексический анализ слова, поиск синонимов, поиск внутренних связей) |
| **Оценка числа многозначных-однозначных слов** | Большинство слов в словарях однозначные, но есть слова со множеством смыслов. Задача состоит в том чтобы подсчитать количество значений слов в электронном словаре и количество многозначных слов |
| **Автоматическое извлечение отношений гипоним-гипероним** | Задача в нахождении всех пар [гипоним-гипероним](https://ru.wikipedia.org/wiki/%D0%93%D0%B8%D0%BF%D0%BE%D0%BD%D0%B8%D0%BC_%D0%B8_%D0%B3%D0%B8%D0%BF%D0%B5%D1%80%D0%BE%D0%BD%D0%B8%D0%BC) для данного словаря. Так как слова многозначны, любое слово может иметь несколько гиперонимов, поэтому если взять все пары [гипоним-гипероним](https://ru.wikipedia.org/wiki/%D0%93%D0%B8%D0%BF%D0%BE%D0%BD%D0%B8%D0%BC_%D0%B8_%D0%B3%D0%B8%D0%BF%D0%B5%D1%80%D0%BE%D0%BD%D0%B8%D0%BC) в словаре в виде группы, то по определению Амслера это будет “запутанная иерархия” |
| **Отображение значений при извлечении сразу из нескольких словарей** | задача состоит в том чтобы сопоставить значения слова, приведенные в одном словаре, со значениями этого же слова в другом словаре (аналог задачи [отображение онтологий](https://ru.wikipedia.org/wiki/%D0%9E%D1%82%D0%BE%D0%B1%D1%80%D0%B0%D0%B6%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BE%D0%BD%D1%82%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D0%B9)). Решение проблемы отображения значений необходимо для того, чтобы объединить данные разных словарях. Например, это позволило бы соотнести между собой грамматическую информацию от Longman Dictionary of Contemporary English с определением из Седьмого Энциклопедического словаря Вебстера |
| **Определение значения в пределах одного словаря** | Отношение синонимии (гиперонимии и др.) — это отношение между значениями, а не словами. Поэтому задача заключается в выборе значения многозначного слова в пределах одного словаря с целью указания синонимичного значения. Эта задача тесно связана с предыдущей, так как при сопоставлении значений нескольких словарей необходимо автоматически выбрать соответствующее значение из списка всех значений, перечисленных в словарной статье |
| **Извлечение информации с помощью набора одноязычных и переводных словарей** | идея в том, чтобы взять информацию из одноязычного словаря для данного языка и, используя переводной (двуязычный) словарь в качестве устройства передачи, передать информацию одноязычного словаря на второй язык |

**Заключение**

Появление и постоянное усовершенствование электронных словарей очень облегчило поиск неизвестных нам слов.

Компьютерная лингвистика не стоит на месте, а развивается. Известный словарь Дж. Миллера **Wordnet** был и останется образцом целевого использования компьютерных технологий в лингвистике. Он даёт более обширный и легкий доступ к нужной нам информации.

Список литературы:

<https://cyberleninka.ru/article/n/rets-na-kn-kotov-a-a-vvedenie-v-leksikografiyu>

<https://studme.org/81716/literatura/leksikografiya>

<https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%BB%D0%B5%D0%BA%D1%81%D0%B8%D0%BA%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D1%8F>

<https://www.abbyy.com/ru-ru/science/technologies/lexicography/>

<https://studfiles.net/preview/5914091/page:17/>

<https://studfiles.net/preview/4342736/>

<https://moluch.ru/archive/21/2055/>

<https://pandia.ru/text/80/454/41928.php>