МИНОБРНАУКИ РОССИИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

“ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ”

*факультет романо-германской филологии*

*кафедра методики и теории преподавания испанского языка*

Компьютерная лексикография

*Курсовая работа*

студентка \_\_\_\_\_\_\_\_\_ Гришанкова В.И

руководитель \_\_\_\_\_\_\_\_\_\_ Донина О.

Воронеж 2018

Оглавление

[1.1. История компьютерной лексикографии 3](#_Toc534974347)

[1.3 Особенности компьютерной лексикографии 6](#_Toc534974348)

[Электронные словари 7](#_Toc534974349)

[**Онлайн-словари** 8](#_Toc534974350)

[§ Преимущества 11](#_Toc534974351)

[Задачи компьютерной лексикографии 13](#_Toc534974352)

[Перед учёными и программистами при преобразовании бумажных словарей в машиночитаемые словари встает множество проблем: 13](#_Toc534974353)

Введение

**Компью́терная лексикогра́фия** - прикладная научная дисциплина в языкознании, которая изучает методы использования компьютерной техники для составления словарей. Это временная дисциплина периода перехода от ручной и рукописной лексикографической практики к новым безбумажным информационным технологиям.

Компьютерная лексикография представлена совокупностью методов и программных средств обработки текстовой информации для создания словарей. В рамках компьютерной лексикографии разрабатываются компьютерные технологии составления и эксплуатации словарей. Специальные программы — базы данных, 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.

Глава 1

## 1.1. История компьютерной лексикографии

Термин **«Компьютерная лексикография»** был придуман для обозначения области изучения машиночитаемых (электронных) словарей и появился в середине 1960-х годов. Этой дисциплине уделяли мало внимания вплоть до начала 1990-х годов. Термин «машиночитаемый словарь» означает, что данные из словаря (хранящиеся в электронном виде) могут быть обработаны и исследованы с помощью современной вычислительной техники.

Новаторские работы Эвенса и Амслера (1980) послужили толчком для расширения исследований электронных словарей, например, была проведена практическая работа с использованием Седьмого Энциклопедического словаря Вебстера. Стимулом для этих исследований послужило также широкое использование Longman Dictionary of Contemporary English в 1980-х, этот словарь по-прежнему является одним из 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. Используя XML, исследователи получают быстрый доступ к информации, хранящейся в электронном словаре.

1.2. Основные понятия компьютерной лексикографии

1. ***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 и автоматические словари для программ обработки текста. 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.
2. ***Гипертекст*** — это множество текстов со связывающими их отношениями (системой переходов).

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

## 

## 

## 

## 

## 

## 

1.3 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:

1. быстро развивающуюся отрасль компьютерной индустрии;
2. 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.

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 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.

В работе (Неси, 2000) выделяют несколько категорий электронных словарей для изучения языков: интернет-словари, глоссарии для учебных онлайн-курсов, словари на компакт-дисках и карманные электронные словари. Несиперечисляет несколько наиболее известных словарей на CD:

* Collins Cobuild Student’s Dictionary
* Cambridge International Dictionaries
* Оксфордский словарь английского языка

### **Онлайн-словари**

Эпоха Интернета сделала онлайн-словари доступными непосредственно с рабочего стола компьютера, а затем и со смартфона. Скинер в 2013 году отметил: «В список слов, которые чаще всего ищут в онлайн-версии словаря Merriam-Webster, сейчас входят слова „holistic“, „pragmatic“, „caveat“, „esoteric“ и „bourgeois“. Исторически целью лексикографии было разъяснение неизвестных слов читателям. И современные словари успешно с этим справляются.»

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
* Double-Tongued Dictionary (данные добавлены пользователями)
* Free On-line Dictionary of Computing
* LEO (website)
* Logos Dictionary
* Pseudodictionary (только юмористические неологизмы, добавленные пользователями)
* Urban Dictionary (словарь англоязычного сленга)
* WWWJDIC (японский онлайн-словарь)
* Визуальный словарь (для каждого слова строится его понятийное окружение).
* Викисловарь
* Русский ассоциативный тезаурус, полученный на основе психолингвистических экспериментов. Интернет-сервис для работы с базой данных ассоциативного эксперимента на русском языке, проведенного в 1988—1997 гг. Ключевой особенностью веб-версии русского ассоциативного тезауруса является возможность проведения компаративного анализа ассоциаций по полу, возрасту и профессии. Тезаурус содержит свыше 1 млн ассоциаций, более 6 тыс. уникальных стимулов и 100 тыс. реакций от более чем 11 тыс. респондентов.

Глава 2.

§ Возможности компьютерной лексикографии

Компьютерная реализация бумажного словаря сама по себе позволяет преодолеть часть указанных проблем. К новым возможностям электронного словаря относятся:

1. Существенно более изощренные возможности показа содержания словарной статьи, включая возможность частичного показа по разным критериям ( различные "проекции" словаря), разнообразные графические средства, которые не используются в обычных словарях.

2. Использование для доступа к содержанию различных лингвистических технологий, таких как морфологический и синтаксический анализ, полнотекстовый поиск, распознавание и синтез звука и т.п.

С точки зрения пользователя смысл реализации в электронном словаре всех этих технологий состоит в том, что становится возможным быстро получить информацию, которая содержится где-то в недрах словаря и непосредственно отвечает тому запросу, который сформулирован пользователем в удобной для него форме.

При традиционном подходе минимальной единицей доступа является лексема (имя словарной статьи): мы должны прочитать всю статью, чтобы определить, содержится ли в ней ответ на наш запрос. Для таких словарей, как оксфордский, это представляет серьезную проблему. Например, глагол set имеет там 400 только основных значений (и у многих из них имеются подзначения).

Пользователь хотел бы, чтобы словарь максимально локализовал релевантную информацию. При этом речь не идет об автоматическом выборе переводного эквивалента (если мы говорим о переводном словаре). Специфика словарного ответа в том, что он дает весьма разнообразную информацию о слове или словосочетании, а не просто переводное соответствие, предполагает активный выбор пользователя из нескольких возможных хорошо обоснованных альтернатив.

Однако, попытка решить проблему адекватной реакции словаря на запрос неизбежно наталкивается на сопротивление самого словарного материала, перенесенного из бумажного словаря.

Преимущества

|  |  |  |  |
| --- | --- | --- | --- |
| Многофункциональность | Использование средств мультимедиа | Актуальность | Большой объем словарной базы |
| Вариативность в использовании | Универсальность | Удобный поиск | Динамичность |

Противоречие  
Итак, мы видим новое противоречие: между новыми языковыми компьютерными технологиями и старым традиционным словарным содержанием, не позволяющим воспользоваться этими технологиями в полном объеме. Иными словами, новые мехи требуют нового вина!

Источник этого противоречия тоже ясен: словарь представляет собой модель языка, устроенную на совершенно иных принципах, чем те формальные модели, которые лежат в основе этих технологий. И если в области морфологии противоречие еще не очень существенно, то в области синтаксиса и семантики оно становится почти непреодолимым.

Действительно, технология морфологического анализа всего лишь позволяет установить соответствие между исходной формой слова из текста и множеством лексем (словарных входов), для которой такая форма возможна. Синтаксический анализ позволяет сделать то же самое для словосочетаний, являющихся отдельными словарными входами.

Однако, для всех этих технологий само словарное содержание является "непрозрачным", полностью ими игнорируется. Заглянуть "внутрь" словарной статьи позволяет только полнотекстовый поиск. Однако, этот мощный инструмент работает со словарным содержимым как с текстом на естественном языке, что резко ограничивает его возможности.

Первый и очевидный шаг, на который уже идут создатели электронных словарей, это первичная разметка словарной статьи, формализация той внутренней структуры, которая в той или иной мере имеется в хороших бумажных словарях.

В результате полнотекстовый поиск может различать, к примеру, переводы, примеры использования и комментарии, что принципиально усиливает его возможности с точки зрения пользователя.

Однако, все эти меры являются поверхностными. Ясно, что задача состоит в том, чтобы единицей описания было отдельное лексическое значение, и технологии анализа могли бы устанавливать соответствие между исходным запросом и теми лексическими значениями, которые релевантны для этого запроса по синтаксическим и семантическим критериям.

В качестве иллюстрирующего примера можно привести практически любой глагол, принадлежащий ядру языка. Например, глагол "развести" может встретиться в таких контекстах:

разводить руками;

разводить спирт водой;

разводить супругов;

разводить мосты;

разводить баранов;

разводить дерущихся;

разводить пилу;

разводить/разбивать сады

(английские эквиваленты: bring; conduct; part, separate; mix; dissolve; divorce; breed; plant, etc...)

Задача создания такого словарного содержания, которое позволило бы сделать единицей анализа отдельное лексическое значение, а не морфологическую лексему, видится нам наиболее перспективным направлением в компьютерной лексикографии. Ясно, что для ее решения требуется "синхронизация" словарных описаний и формальных моделей, используемых технологиями анализа. В пределе это должно быть единое интегральное лексико-синтактико-семантическое описание.

## 

## 

Задачи компьютерной лексикографии

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

Таблица 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Парсинг** — 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 для данного словаря. Так как слова многозначны, любое слово может иметь несколько гиперонимов, поэтому если взять все пары гипоним-гипероним в словаре в виде группы, то по определению Амслера это будет «запутанная иерархия»(англ. *tangled hierarchy)* | **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). 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 |

Список использованных источников

1. <https://www.abbyy.com/ru-ru/science/technologies/lexicography/>
2. <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>
3. <http://www.wikiwand.com/ru/%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>
4. <https://studfiles.net/preview/5914091/page:17/>

Приложение

* *Баранов А.Н.* [Введение в прикладную лингвистику](http://aska1.ru/files/%D0%91%D0%B0%D1%80%D0%B0%D0%BD%D0%BE%D0%B2%20%D0%90.%D0%9D.%20-%20%D0%92%D0%B2%D0%B5%D0%B4%D0%B5%D0%BD%D0%B8%D0%B5%20%D0%B2%20%D0%BF%D1%80%D0%B8%D0%BA%D0%BB%D0%B0%D0%B4%D0%BD%D1%83%D1%8E%20%D0%BB%D0%B8%D0%BD%D0%B3%D0%B2%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D1%83%20(2001).doc). — Эдиториал УРСС, 2001. — P. 360.
* *Кашеварова И.С.* [Электронный словарь как новый этап в развитии лексикографии](http://www.gendocs.ru/docs/27/26246/conv_1/file1.pdf#page=145). — «Молодой ученый», 2010. — № 10 (21). — P. 145-147.
* *Леонова А.В., Снопкова Н.А.* [СОВРЕМЕННЫЕ ИНСТРУМЕНТЫ ТЕХНИЧЕСКОГО ПЕРЕВОДЧИКА](http://www.istu.edu/docs/science_periodical/mvestnik/Leonova%20.doc). — Национальный исследовательский Иркутский государственный технический университет, 2012. — P. 2.
* *Сивакова Н.А.* [Лексикографическое описание английских и русских фитонимов в электронном глоссарии](http://www.tmnlib.ru/DbFileHandler.axd?376) // Тюмень. — 2004. — P. 28.
* *Соловьева А.В.* [Профессиональный перевод с помощью компьютера](http://www.google.ru/books?hl=ru&lr=&id=jAGO9QcMaEMC&oi=fnd&pg=PA3&dq=%D0%9F%D0%BE%D1%80%D1%82%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D1%8B%D0%B5+%D1%8D%D0%BB%D0%B5%D0%BA%D1%82%D1%80%D0%BE%D0%BD%D0%BD%D1%8B%D0%B5+%D1%81%D0%BB%D0%BE%D0%B2%D0%B0%D1%80%D0%B8&ots=1Rn5JnrVne&sig=knlu3dbx2Z7Qxq0nMsT_wU1drGk&redir_esc=y#v=onepage&q=%D0%9F%D0%BE%D1%80%D1%82%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D1%8B%D0%B5%20%D1%8D%D0%BB%D0%B5%D0%BA%D1%82%D1%80%D0%BE%D0%BD%D0%BD%D1%8B%D0%B5%20%D1%81%D0%BB%D0%BE%D0%B2%D0%B0%D1%80%D0%B8&f=false). — Питер, 2008.
* *Табанакова В.Д., Ковязина М.А.* [Новая модель двуязычного экологического словаря-тезауруса](http://tc.utmn.ru/files/kovyazina_tabanakova_ecothesaurus.pdf). — Урал. гос. пед. ун-т, 2007. — P. 29 – 32. (недоступная ссылка)
* *Филиппович Ю., Чернышева М.* [Историческая компьютерная лексикография - terra incognita в компьютерном мире](http://old.computerra.ru/1999/323/3379/). — Компьютерра, 1999. — № 45 (09 ноября 1999). (недоступная ссылка)
* *Чепик Е.Ю.* [Компьютерная лексикография как одно из направлений современной прикладной лингвистики](http://www.science.crimea.edu/zapiski/2006/filologiya/uch_19_2fn/chepik_51.pdf). — Ученые записки таврического национального университета им В.И. Вернадского., 2006. — Т. 19, № 3-4. — P. 274-279. [Архивировано](https://web.archive.org/web/20140420072214/http://www.science.crimea.edu/zapiski/2006/filologiya/uch_19_2fn/chepik_51.pdf) 20 апреля 2014 года.
* *Чепик Е.Ю.* [Политическое слово в структуре электронного словаря.](http://dspace.nbuv.gov.ua/bitstream/handle/123456789/10641/57-Chepik.pdf?sequence=1). — Культура народов Причерноморья., 2005. — № 69. — P. 205-209.
* *Чепик Е.Ю.* [Политическое слово в структуре электронного словаря.](http://dspace.nbuv.gov.ua/bitstream/handle/123456789/10641/57-Chepik.pdf?sequence=1). — Культура народов Причерноморья., 2005. — № 69. — P. 205-209.
* *Шемякин Ю.И.* [Начала компьютерной лингвистики](http://www.ict.edu.ru/ft/004938/shemakin.pdf). — Издательство МГОУ А/О "Росвузнаука", 2003. — P. 81.