Учреждение образования
«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ»

Кафедра интеллектуальных информационных технологий

**Лабораторная работа №3 по курсу «ЕЯзИИС» на тему:**
«Представление результатов синтаксического анализа в памяти интеллектуальной системы»
Вариант 8

Выполнили студенты группы 021701: Кулак П.О., Седеневский А.М., Малаев А.А.

Проверил: Крапивин Ю.Б.

**МИНСК**
2023

**Цель:**

Закрепить знания семантического анализа текста.

**Основные задачи:**

1. Изучить теоретический материал, необходимый для решения задачи автоматического семантического анализа текста естественного языка.

2. Закрепить навыки программирования естественно-языковых систем и вспомогательных прикладных программ на одном из языков программирования.

**Вариант:**

| № Варианта | Язык текста | Функциональность | Формат документа |
|---|---|---|---|
| 8 | Английский | WordNet | txt |

**Порядок выполнения работы:**
В ходе выполнения лабораторной работы была написана программа для семантического анализа предложения подаваемого на вход или из txt-файла, выбранного пользователем.

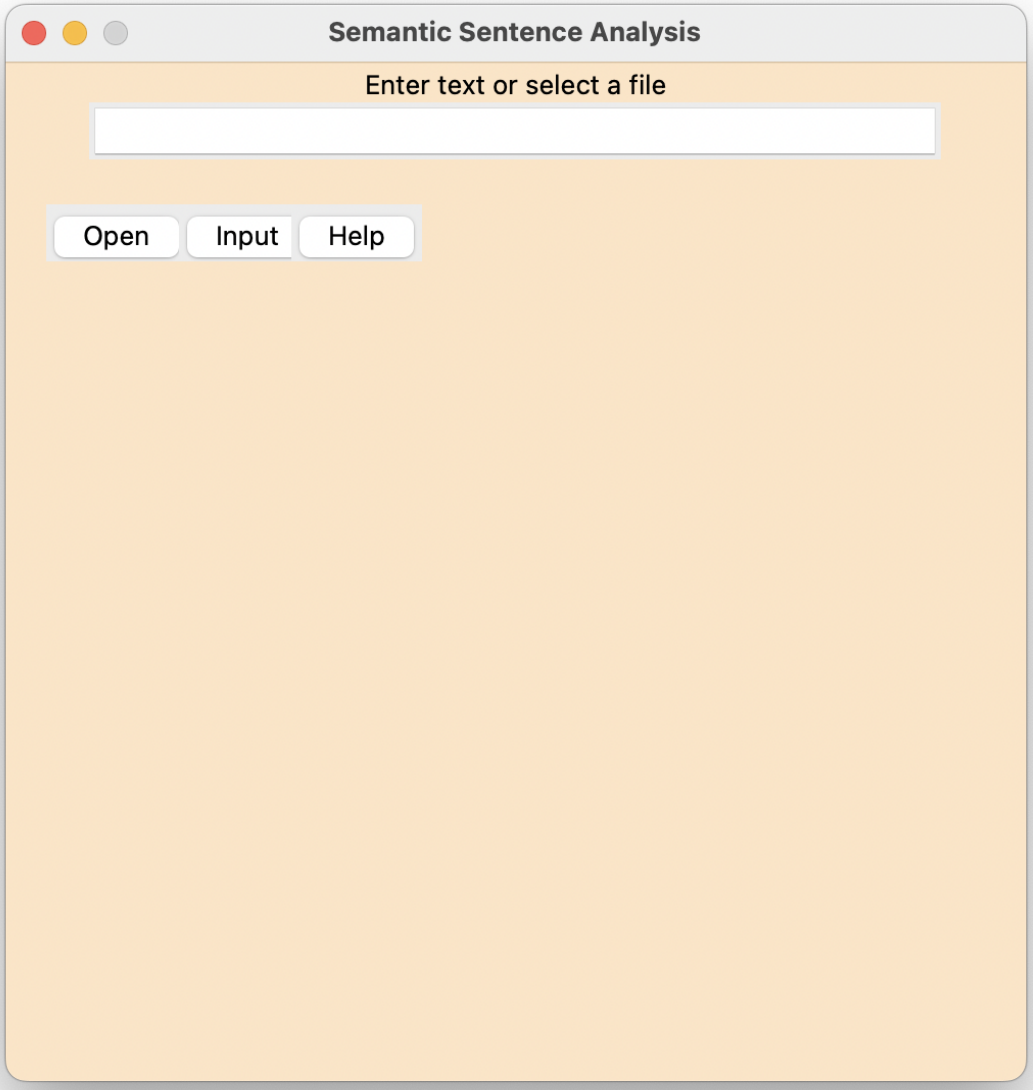


*Рисунок 1 Внешний вид программы*

Внешний интерфейс программы состоит из строки ввода и 3 кнопок:

- Кнопка Open позволяет выбрать файл, содержимое которого будет анализироваться
- Кнопка Input позволяет проанализировать текст введённый в строку
- Кнопка Help для помощи пользователю разобраться с ходом работы программы

**Для создания данного приложения былиспользован леммитатор WordNet из библиотеки NLTK.**

NLTK - ведущая платформа для создания программ Python для работы с данными на человеческом языке. Он предоставляет простые в использовании интерфейсы для более чем 50 корпусных и лексических ресурсов, таких как WordNet, а также набор библиотек обработки текста для классификации, токенизации, стемминга, тегирования, синтаксического анализа и семантического анализа, оболочки для промышленных библиотек NLP, и активный дискуссионный форум.

Благодаря практическому руководству, представляющему основы программирования наряду с темами вычислительной лингвистики, а также исчерпывающей документации по API, NLTK подходит как для лингвистов, инженеров, студентов, преподавателей, исследователей, так и для пользователей отрасли. NLTK доступен для Windows, Mac OS X и Linux. Лучше всего то, что NLTK - это бесплатный проект с открытым исходным кодом, управляемый сообществом.
NLTK был назван «прекрасным инструментом для обучения и работы в области компьютерной лингвистики с использованием Python» и «потрясающей библиотекой для игры с естественным языком».

Wordnet — это большая, свободно распространяемая и общедоступная лексическая база данных для английского языка с целью установления структурированных семантических отношений между словами. Библиотека также предлагает возможности лемматизации и является одним из самых ранних и наиболее часто используемых лемматизаторов.

В ходе работы были использованы следующие методы:

```
Syn.examples() – возвращает примеры использования слова в тексте
Syn.definition() – возвращает определение слова
Syn.antonyms()- возвращает антонимы слова
Syn.hypernyms()- возвращает гипернимы слова(слово с более широким
значением, выражающее общее, родовое понятие, название класса
предметов)
Syn.hyponyms() –возвращает гипонимы слова(понятие, выражающее частную
сущность по отношению к другому, более общему понятию)
```

**Тестирование системы:**
Для тестирования работы приложения было введено слово ‘angry’ и запущен анализ строки.

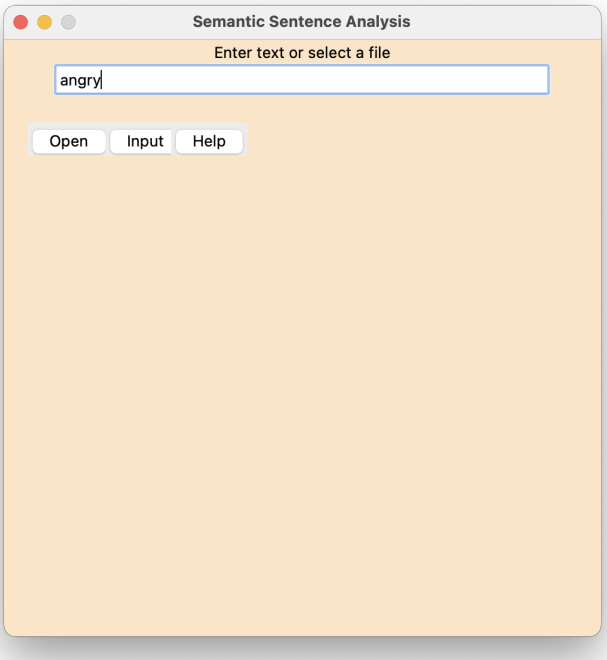


*Рисунок 2 Входные данные*

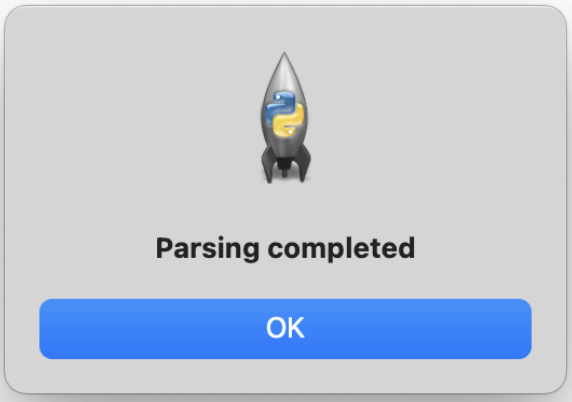


*Рисунок 3 Ответ программы*

Результат анализа был сохранён в файл semantic_analysis.txt

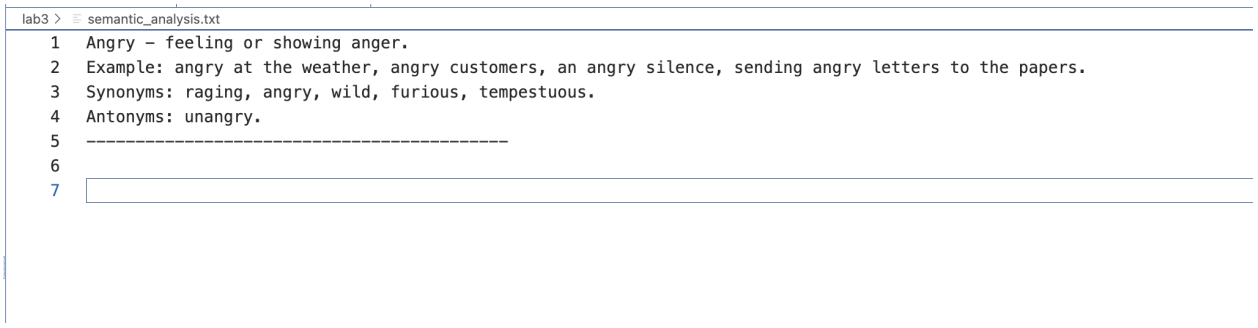


*Рисунок 4 Результат работы программы*

Далее был проведён тест работы программы выбором файла с расширением txt.

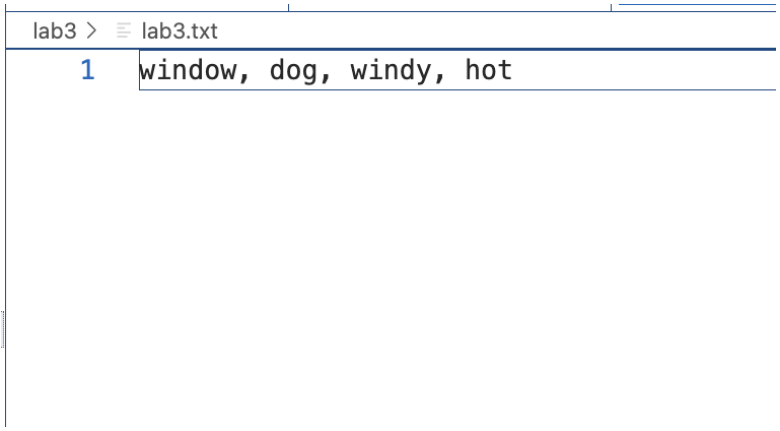


*Рисунок 5 Содержимое файла*

ХЧерез приложение был выбран данный файл

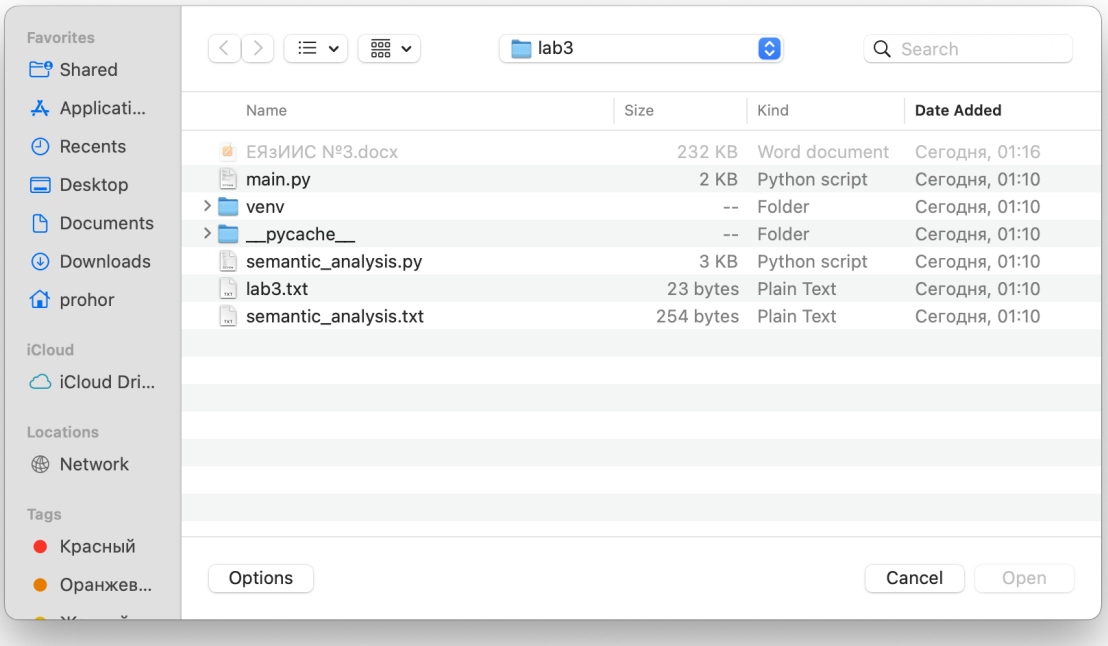


*Рисунок 6 Выбор файла*

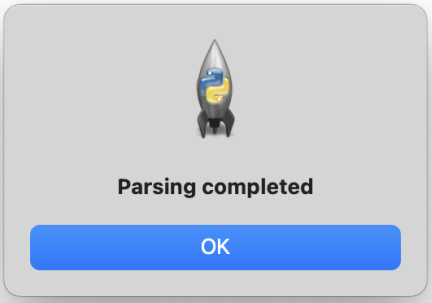


*Рисунок 7 Ответ программы*

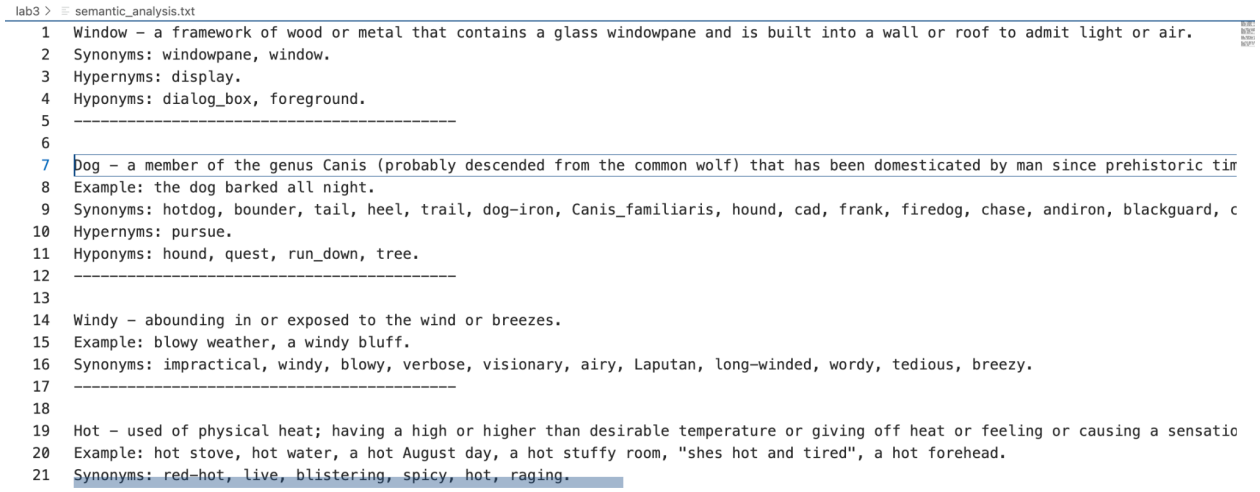


*Рисунок 8 Результат работы программы*

**Вывод**

В ходе лабораторной работы было создано приложение для семантического анализа текста на английском языке. Так же были приобретены навыки работы с лексичной базой WordNet. В систему была внедрена помощь пользователю, а также протестирована. Скорость обработки текста длиной в одно слово составляет примерно 1.84 секунды, обработки текста длиной в 4 слова - 1.94. Основное время работы программы тратится на загрузку базы WordNet.

**9. Автоматизация обработки текста: этап лексико-грамматического анализа.**

Лексико-грамматический анализ (лемматизация и морфологический анализ) является одним из этапов автоматизации обработки текста на естественном языке. Его задача заключается в определении лексических и грамматических свойств каждого слова в предложении.

Лемматизация - это процесс приведения слова к его базовой форме (лемме), которая может использоваться для поиска похожих слов и выявления общих понятий. Например, слова "бегать", "бегу", "бегает" будут приведены к лемме "бегать".

Морфологический анализ - это процесс определения грамматических свойств каждого слова в предложении, таких как число, род, падеж, время и т.д. Эти свойства помогают понимать, как слова соотносятся друг с другом в предложении.

Для лексико-грамматического анализа используются различные методы, включая статистические модели, правила и машинное обучение. Он может быть выполнен как на основе словарей и правил, так и на основе машинного обучения, используя нейронные сети и другие алгоритмы.

Цель лексико-грамматического анализа - предоставить программе информацию о структуре и значении текста, что позволяет дальнейшим этапам обработки более точно и корректно анализировать текст.

Достоинством лексико-грамматического анализа является то, что он позволяет программе понимать грамматическую структуру предложения и использовать эту информацию для более точного анализа текста. Кроме того, этот этап позволяет обнаруживать и исправлять ошибки в написании слов.

Недостатком лексико-грамматического анализа может быть сложность обработки некоторых специфичных языковых конструкций, таких как сленговые и диалектные выражения, что может приводить к неточностям в анализе.

**10. Модель СМЫСЛ – ТЕКСТ: определение, компоненты, задачи.**

Модель СМЫСЛ - ТЕКСТ (Semantic Text Model) - это модель, которая описывает процесс формирования смысла текста на основе его лексических и синтаксических структур. Она состоит из двух компонентов - компонента смысла и компонента текста.

Компонент смысла определяет семантику текста, т.е. его смысловую нагрузку. Он включает в себя лексический, семантический и прагматический анализ текста. Лексический анализ определяет значения слов, семантический анализ определяет отношения между словами в предложении, а прагматический анализ определяет намерения и цели автора.

Компонент текста определяет структуру текста и его связи. Он включает в себя синтаксический и лексический анализ. Синтаксический анализ определяет связи между словами в предложении, а лексический анализ определяет значения слов и выражений в контексте.

Задачи модели СМЫСЛ - ТЕКСТ включают:

- Автоматическое резюмирование текста: создание краткого описания содержания текста.
- Автоматический перевод: перевод текста с одного языка на другой.
- Классификация текста: определение жанра, темы или других характеристик текста.
- Анализ тональности: определение эмоциональной окраски текста.
- Извлечение информации: выделение значимой информации из текста.

Одним из главных достоинств модели СМЫСЛ - ТЕКСТ является ее способность понимать контекст и связи между словами в предложении, что позволяет более точно и эффективно анализировать текст. Однако недостатком этой модели может быть сложность ее реализации из-за сложности лексических и синтаксических структур естественного языка.

**11. Структура диалоговой системы.**

Диалоговая система - это программа, которая может взаимодействовать с пользователем на естественном языке. Структура диалоговой системы включает в себя несколько основных компонентов:

1. Распознавание речи - этот компонент отвечает за преобразование звуковой волны речи пользователя в текстовую форму, которую может обработать дальнейший компонент системы.
2. Понимание естественного языка - этот компонент анализирует текст, полученный от распознавания речи, и пытается понять, что пользователь хочет сказать. Он определяет намерение пользователя и извлекает информацию из текста.
3. Генерация ответа - на основе информации, извлеченной из текста, система формулирует ответ на вопрос или запрос пользователя. Этот ответ может быть сгенерирован автоматически на основе заранее подготовленных ответов или же система может обращаться к базе знаний, чтобы получить необходимую информацию.
4. Синтез речи - этот компонент преобразует текстовый ответ, сформулированный системой, в звуковую волну, которую пользователь может услышать.
5. Управление диалогом - этот компонент отвечает за управление потоком диалога между пользователем и системой, например, задает вопросы пользователю, чтобы уточнить информацию, или предлагает пользователю некоторые варианты ответов.

6. База знаний - это хранилище информации, которое система использует для ответов на вопросы пользователей. Она может содержать информацию о товарах, услугах, расписании работы и т.д.
7. Интерфейс пользователя - это часть системы, которая обеспечивает взаимодействие между пользователем и системой. Это может быть голосовой ассистент, приложение на мобильном устройстве или сайт с возможностью чата.

Структура диалоговой системы может варьироваться в зависимости от ее целей и функций, но в целом она должна включать все вышеперечисленные компоненты.

**12. Лингвистический процессор.**

Лингвистический процессор - это компьютерная программа, которая проводит автоматический лингвистический анализ текста на естественном языке. Он состоит из нескольких компонентов, каждый из которых выполняет определенную функцию, такую как лексический, морфологический, синтаксический и семантический анализ. Лингвистический процессор может использоваться для различных целей, таких как автоматический перевод, извлечение информации, анализ тональности, генерация текста и другие. Он может работать как в режиме онлайн-обработки, так и в режиме обработки больших объемов текста в автономном режиме.

Компоненты лингвистического процессора:

1. Лексический анализатор - отвечает за разбор текста на лексические единицы, такие как слова, числа, знаки пунктуации, идентификацию неизвестных слов и т.д.
2. Морфологический анализатор - определяет грамматическую форму слова и его часть речи.
3. Синтаксический анализатор - анализирует синтаксическую структуру предложения и определяет зависимости между словами в предложении.
4. Семантический анализатор - анализирует смысл текста и связывает его с знаниями, хранящимися в базе знаний.
5. Генератор текста - создает новый текст на основе полученных результатов анализа.
6. Интерфейс пользователя - это часть системы, которая обеспечивает взаимодействие между пользователем и лингвистическим процессором. Это может быть голосовой ассистент, приложение на мобильном устройстве или сайт с возможностью чата.

Каждый из компонентов лингвистического процессора выполняет свою задачу, и результаты анализа используются для решения различных задач в области обработки естественного языка.

5. Прикладная лингвистика - это область знаний, которая занимается применением теоретических знаний лингвистики к практическим проблемам, связанным с использованием языка в реальных ситуациях. Она занимается анализом, описанием и практическим применением языковых знаний в различных областях деятельности.

Компьютерная лингвистика и прикладная лингвистика тесно связаны друг с другом, но имеют и некоторые различия:

- Компьютерная лингвистика фокусируется на разработке и применении технологий и методов обработки естественного языка, тогда как прикладная лингвистика применяет

лингвистические знания для решения конкретных языковых проблем в различных сферах деятельности, таких как образование, межкультурная коммуникация, перевод и т.д.

- Компьютерная лингвистика использует технологии и методы искусственного интеллекта, машинного обучения и компьютерных алгоритмов для обработки естественного языка, тогда как прикладная лингвистика обычно использует более традиционные методы и технологии для решения языковых проблем.

- В области компьютерной лингвистики уделяется больше внимания разработке и улучшению технологий для автоматического анализа и генерации естественного языка, тогда как прикладная лингвистика более ориентирована на практические задачи, такие как создание учебных
материалов для изучения языка, обеспечение эффективной коммуникации на рабочем месте и т.д.

Таким образом, прикладная лингвистика и компьютерная лингвистика имеют много общих точек соприкосновения, но каждая из них имеет свои специфические задачи и подходы к решению языковых проблем.

6. Компьютерная лингвистика возникла из необходимости обработки больших объемов текстовой информации и автоматической обработки естественного языка. Возникновение компьютерной лингвистики обусловлено следующими причинами:

- Развитие компьютерных технологий: С появлением компьютеров и программных средств для обработки текстов, стало возможным создание автоматических систем обработки естественного языка.

- Увеличение объема текстовой информации: С ростом объема текстовой информации стало необходимым разрабатывать методы автоматической обработки и анализа этой информации.

- Появление новых форм коммуникации: С развитием интернета и социальных сетей появились новые формы коммуникации, которые требуют автоматической обработки естественного языка, такие как обработка сообщений и комментариев в социальных сетях.

- Необходимость автоматизации процессов: В различных сферах деятельности, таких как банковское дело, медицина, право, наука и техника, стала необходима автоматизация процессов, в том числе и обработки текстовой информации.

- Развитие искусственного интеллекта: Компьютерная лингвистика является частью области искусственного интеллекта, которая занимается разработкой и применением методов и технологий для создания умных систем и решения сложных задач.

Таким образом, компьютерная лингвистика возникла из необходимости решения практических проблем, связанных с обработкой естественного языка, и является важной областью в области информационных технологий.