1. **Лингвистическая защита информации** (ЛЗИ) — это защита информации на этапе опредмечивания, овеществления смыслов, когда адресант, апеллируя к адресату, воплощает эти смыслы в сообщении. **Для защиты персональных данных лингвисты могут использовать следующие методы:**

* Анализ идентификационной информации.
* Анализ контекста.
* Разработка политик безопасности.
* Усовершенствование онлайн-инструментов для защиты персональных данных.
* Разработка правильной лингвистической тактики в социальных сетях и онлайн-мессенджерах.

1. Прикладная лингвистика не является синонимом терминов «компьютерная лингвистика», «вычислительная лингвистика» или «инженерная лингвистика», хотя в зарубежных исследованиях под термином «computational linguistics» понимаются все эти сферы и в научной литературе не встречаются отдельные термины для каждого из этих направлений. Нужно понимать, что прикладная лингвистика - это не использование компьютерных программ, это решение проблем, связанных с языком, это «приложение» любых знаний и технологий для решения практических задач.
2. Основными направлениями, рассматриваемыми в курсе компьютерной лингвистики, являются: информационный поиск, машинный перевод, терминология, терминоведение, терминография, компьютерная лексикография, распознавание и синтез речи, проблемы обучения языку с помощью компьютера. К компьютерной лингвистике относят все, что связано с использованием компьютеров в языкознании. Машинный перевод (Machine Translation) [20] — самое раннее приложение КЛ, вместе с которым возникла и развивалась сама эта область. Первые программы перевода были построены в середине прошлого века и были основаны на простейшей стратегии пословного перевода. Однако довольно быстро было осознано, что машинный перевод требует гораздо более полной лингвистической модели. Такая модель была разработана в отечественной системе ЭТАП [24], а также в нескольких других системах, выполняющих перевод научных текстов. В настоящее время существует целый спектр компьютерных систем машинного перевода (разного качества), от больших интернациональных исследовательских проектов до коммерческих автоматических переводчиков. Существенный интерес представляют проекты многоязыкового перевода с использованием промежуточного языка, на котором кодируется смысл переводимых фраз.

Упражнения: Уровень 1

Раздел 1:

|  |  |
| --- | --- |
| Слов | 67 |
| Символов (без пробелов) | 539 |
| Символов (с пробелами) | 605 |
| Символов в латинской графике | 95 |
| Чисел | 3 |
| Средняя длина слов | 8,04 |

1. Средняя длина слова составляет около 6 букв, но зависит от корпуса (набора) текстов. Для корпуса русской разговорной речи средняя длина слова обычно составит от 3.9 до 4.9 букв. Для корпуса русской художественной литературы - от 4.9 до 5.9. Для газетно-новостного-корпуса - от 5.9 до 6.9. И наконец, для корпуса научной и деловой литературы - от 6.9. Средняя длина английского слова составляет пять букв. Следовательно, средняя длина слов текста смешанного типа, представленного в задании 1, соответствует русскому языку.
2. а) Ø – датский, (ö ß) – немецкий, ρ ω – греческий, ё – русский, š – эстонский; б) th – английский, sch – немецкий, šč – словенский; в) et – французский, the – английский, der – немецкий, och – шведский, için – турецкий.

|  |  |  |
| --- | --- | --- |
| Критерий | Язык 1: Английский | Язык 2: Испанский |
| Типичные артикли | the, a, an | el, la, los, las |
| Указательные местоимения | this, that, these, those | este, esta, estos, estas, ese, esa, esos, esas |
| Местоимения 3-го лица | he, she, it | él, ella, ello |
| Отдельные формы вспомогательных глаголов | do, does, did; will, would; can, could | haber (he, has, ha, hemos, han); ser (soy, eres, es, somos, son); estar (estoy, estás, está, estamos, están) |
| Основные предлоги и союзы | in, on, at; and, but, or | en, a; y, pero |
| Другие частотные слова | to, is, are; not, can; I, you; he, she; we, they | de, que; no; yo, tú; él, ella; nosotros/nosotras, ellos/ellas |

1. Для дополнения таблицы встречаемости букв в распространенных европейских языках добавлю данные по русскому языку на основе следующего текста: "Русский язык - один из самых красивых и богатых языков мира. Он обладает богатой историей, разнообразным словарным запасом и глубокими корнями. В русском языке используется кириллический алфавит, который состоит из 33 букв. Среди самых часто встречающихся букв в русском языке можно выделить буквы: о, а, е, и, н, т, с, р, в, л." Теперь дополненная таблица встречаемости букв в распространенных европейских языках выглядит следующим образом:

|  |  |  |  |
| --- | --- | --- | --- |
| Буква | Английский язык | Испанский язык | Русский язык |
| А | 8,2% | 12,5% | 7,6% |
| Б | 1,5% | 0,9% | 1,6% |
| В | 3,3% | 1,4% | 4,3% |
| Г | 1,9% | 1,1% | 1,7% |
| Д | 4,2% | 3,1% | 2,9% |
| Е | 12,7% | 14,7% | 8,4% |
| Ж | 0,6% | 0,4% | 0,8% |
| З | 1,7% | 0,7% | 1,6% |
| И | 7,5% | 7,5% | 6,2% |
| Й | 1,5% | 0,5% | 1,2% |
| К | 6,6% | 3,3% | 3,5% |
| Л | 4,3% | 4,6% | 3,2% |
| М | 2,5% | 3,1% | 2,9% |
| Н | 7,2% | 7,5% | 6,7% |
| О | 7,6% | 5,4% | 10,9% |
| П | 2,0% | 3,1% | 2,5% |
| Р | 6,7% | 6,6% | 4,7% |
| С | 6,7% | 8,0% | 5,9% |
| Т | 9,1% | 7,3% | 6,8% |
| У | 2,8% | 4,4% | 2,6% |
| Ф | 0,2% | 0,9% | 0,3% |
| Х | 1,5% | 0,8% | 0,9% |
| Ц | 0,6% | 1,4% | 0,5% |
| Ч | 1,2% | 0,8% | 1,6% |
| Ш | 0,9% | - | 0,8% |
| Щ | 0,3$ | - | - |
| Ъ | - | - | - |
| Ы | - | - | - |
| Ь | - | - | - |
| Э | - | - | - |
| Ю | - | - | - |
| Я | - | - | - |

Теперь таблица содержит данные по встречаемости букв в английском, испанском и русском языках для сравнения и анализа.

1. Морфологические диагностические показатели эсперанто: 1. **Суффиксы**: В эсперанто существительные и прилагательные имеют одинаковые окончания. Например, суффикс "-a" обозначает прилагательное (verd**a** - зелен**ый**), а суффикс "-o" обозначает существительное (arb**o** - дерев**о**). 2. **Грамматическое родоведение**: В эсперанто нет грамматического рода у существительных и прилагательных. Они не изменяются по роду, что упрощает грамматическую структуру языка. 3. **Единая форма для разных частей речи**: В эсперанто одинаковые окончания используются для образования различных частей речи, что делает язык более предсказуемым и легким для изучения. 4. **Словообразование**: В эсперанто слова образуются посредством аффиксации (добавление приставок и суффиксов) к корням, что делает процесс формирования новых слов структурированным и прозрачным.

Раздел 2:

1. С точки зрения компьютерной семантики и компьютерной грамматики можно рассмотреть данную задачу следующим образом: 1. **Ландыш** - это цветок. 2. **Левкой** - это также цветок. 3. **Лаватера** - это цветок. 4. **Лютик** - это цветок. 5. **Люпин** - это цветок. 6. **Ромашка** - это цветок. 7. **Липа** - это дерево. Слово **Липа** является лишним в данном списке, так как все остальные слова представляют собой названия цветов, а "Липа" - это дерево.
2. а) AnNn ↔ N(A)nNg (быстрое движение ↔ быстрота движения) - An - прилагательное, обозначающее характеристику или свойство - Nn - существительное, обозначающее объект или явление - Ng - существительное, обозначающее абстрактное понятие - N(A)n - существительное, обозначающее объект или явление, к которому относится прилагательное - Пример цитаты: "быстрое движение ↔ быстрота движения" b) VNa ↔ N(V)nNg (прибавляю число ↔ прибавление числа) - V - глагол, обозначающий действие - Na - существительное, обозначающее объект или явление - N(V)n - существительное, обозначающее объект или явление, к которому относится глагол - Ng - существительное, обозначающее абстрактное понятие - Пример цитаты: "прибавляю число ↔ прибавление числа" c) N1nN2g ↔ A(N2)nN1n (права автора ↔ авторские права) - N1 - существительное, обозначающее объект или явление - N2 - существительное, обозначающее объект или явление - Ng - существительное, обозначающее абстрактное понятие - A(N2)n - прилагательное, обозначающее характеристику или свойство, к которому относится существительное - Пример цитаты: "права автора ↔ авторские права" d) VвNa ↔ N(V)nвNa (возвожу в степень ↔ возведение в степень) - V - глагол, обозначающий действие - Na - существительное, обозначающее объект или явление - N(V)n - существительное, обозначающее объект или явление, к которому относится глагол - в - предлог - Пример цитаты: "возвожу в степень ↔ возведение в степень" e) DV ↔ A(D)nN(V)n (сильно желать ↔ сильное желание) - D - наречие, обозначающее степень действия - A - прилагательное, обозначающее характеристику или свойство - (D)n - наречие, обозначающее степень действия, к которому относится прилагательное - N(V)n - существительное, обозначающее объект или явление, к которому относится глагол - Пример цитаты: "сильно желать ↔ сильное желание"
3. а) AnNn ↔ N(A)nNg - визит врача ↔ врачебный визит - Пример цитаты: "визит врача ↔ врачебный визит" б) N1nN2g ↔ A(N2)nN1n - ароматный сад ↔ аромат сада - Пример цитаты: "ароматный сад ↔ аромат сада" в) VNa ↔ N(V)nNg - выхожу из дома ↔ выход из дома - Пример цитаты: "выхожу из дома ↔ выход из дома"
4. а) книга из библиотеки ↔ библиотечная книга. Пример цитаты: "Я взял книгу из библиотеки и узнал, что это библиотечная книга." б) зеленая трава шумно шелестит под легким ветром.

Уровень 2

1. Этапы развития компьютерной лингвистики: 1. 1950-е годы: Ранние работы в области машинного перевода. 2. 1960-е годы: Развитие области компьютеризованной лексикографии. Появление искусственного интеллекта и робототехники. 3. 1970-е годы: Исследования в области компьютерной обработки естественного языка (NLP). Развитие лингвистической базы данных. 4. 1980-е годы: Возникновение эффективных алгоритмов компьютерного распознавания и синтеза речи. Появление методов машинного обучения в NLP. 5. 1990-е годы: Расцвет интернет-технологий и развитие текстовых аналитических инструментов. Переход от статистических методов к глубокому обучению в NLP. 6. 2000-е годы и последующие: Использование больших данных (Big Data) и облачных вычислений в NLP. Развитие прикладных областей, таких как машинный перевод, автоматическая обработка текста и сентимент-анализ.

Пути развития компьютерной лингвистики: 1. Глубокое обучение и искусственный интеллект: Применение глубоких нейронных сетей для более точной обработки текста и речи. 2. Мультиязычные системы и межязыковая адаптация: Разработка универсальных алгоритмов для работы с различными языками и перевода между ними. 3. Интеграция с другими областями: Совмещение компьютерной лингвистики с информационными технологиями, биоинформатикой и другими областями для синтеза новых методов и приложений. 4. Развитие автоматизированных систем NLP: Создание более эффективных и интуитивных систем обработки текста для конечных пользователей. 5. Эксплорация области обучения с подкреплением: Исследование методов обучения с подкреплением в контексте компьютерной лингвистики для достижения лучшей производительности.

1. Потенциальные причины сбоя: Некачественные аудиозаписи или шум на фоне могут затруднить распознавание речи. Недостаточное обучение алгоритмов на различных акцентах или интонациях может привести к ошибкам в распознавании. Ограниченные вычислительные ресурсы могут привести к задержкам или ошибкам при обработке аудио данных. Сбои в работе компьютерных программ, особенно в области распознавания и анализа речи, могут иметь множество причин, связанных как с алгоритмами и моделями машинного обучения, так и с техническими аспектами, такими как качество аудио данных, соединение и вычислительные мощности.
2. (Древние греки придумали Олимпийские игры, пока вели одну из своих нескончаемых войн. Основных причин было две: во-первых, во время баталий солдатам и офицерам некогда было заниматься спортом, а ведь эллины (так называли себя древние греки) стремились тренироваться всё время, не занятое упражнениями в философии; во-вторых, воинам хотелось поскорее вернуться домой, а отпуск на войне не предоставлялся. Было ясно, что войска нуждались в перемирии и что единственной возможностью его объявить могли стать Олимпийские игры, ведь непременное условие Олимпиады — прекращение войны. За множество веков Олимпиада преодолела не одно препятствие. Сейчас это одно из крупнейших и захватывающих событий в мире, во время которого по-прежнему стараются прекратить все военные конфликты.)

Частотный анализ поможет определить, какие слова встречаются чаще всего в тексте и как они распределены. 1. "войны" - 3 раза 2. "Олимпийские" - 3 раза 3. "игры" - 3 раза 4. "время" - 2 раза 5. "воинам" - 2 раза 6. "войска" - 1 раз 7. "прекращение" - 1 раз 8. "препятствие" - 1 раз 9. "крупнейших" - 1 раз 10. "событий" - 1 раз

1. Квантитативная лингвистика - это подход в лингвистике, который использует количественные методы для изучения языка, а не только качественные аспекты. Она занимается изучением языковых явлений с помощью статистики, математики и данных. Компьютерная лингвистика - это область, которая объединяет лингвистику и информатику для решения задач обработки естественного языка, таких как машинный перевод, распознавание и синтез речи, анализ текста и другие. Компьютерная лексикография - это раздел лингвистики, который занимается созданием и использованием электронных словарей с помощью компьютерных технологий. Отличие компьютерной лексикографии от традиционной связано с использованием современных методов обработки текста, баз данных и интерактивных функций для создания словарей. Среди современных методов анализа текста можно выделить такие подходы как: - Машинное обучение и нейронные сети для анализа текста. - Обработка естественного языка (NLP) для извлечения информации и анализа текстов. - Сентимент-анализ для определения эмоциональной окраски текста.
2. Ментальная карта квантитативной лингвистики - Определение - Квантитативная лингвистика: использование статистических методов и математики для анализа языка. - Методы и техники анализа - Статистический анализ текстов - Анализ корпусов языка - Частотный анализ слов - Использование компьютерных программ для обработки данных - Языковые модели и представления - N-граммы и марковские модели - Векторные представления слов - Машинное обучение для языковых задач - Приложения квантитативной лингвистики - Машинный перевод - Автоматическое извлечение информации - Анализ сентиментов в тексте - Классификация текстов - Распознавание речи - Инструменты и программное обеспечение - Python и R для анализа данных - Библиотеки для обработки текста, такие как NLTK и SpaCy - Инструменты для создания и анализа корпусов текстов - Исследовательские направления в квантитативной лингвистике - Статистический анализ языка - Когнитивная лингвистика - Анализ социальных медиа и интернет-данных
3. Первый представительный корпус английского языка под названием BNC (British National Corpus). Корпус включал тексты различных жанров, в том числе художественную литературу, научные статьи, периодическую печать и разговорную речь. BNC стал важным инструментом для исследования английского языка и использовался как основа для создания словарей и разработки языковых моделей. В его создании участвовали лингвисты, компьютерные специалисты и издатели. Корпус был доступен для всех заинтересованных исследователей и стал одним из основных компонентов в лингвистических исследованиях. Разметки корпусов является xces. Он предоставляет определенные элементы и атрибуты для разметки различных лингвистических признаков, таких как лемма, часть речи, грамматические категории и др. XCES также предлагает возможность создания словарей, описывающих значения лингвистических признаков. Это позволяет не только разметить корпус, но и проводить более глубокий анализ и исследования на основе семантической и синтаксической информации.\n\nТаким образом, разметка корпуса играет важную роль в его использовании и анализе. Разные корпуса могут быть размечены разными методами и с использованием различных стандартов, в зависимости от задач и целей исследования. Корпусы используются в лингвистических исследованиях для анализа различных языковых явлений. Компьютерные корпуса не меняют основную деятельность лингвиста, но предоставляют новые возможности и инструменты для исследований. Создание представительных корпусов и их разметка на различных уровнях позволяют строить словари и грамматики, основанные на корпусной методологии. Например, создание корпуса LOB (построенного на модели Brown-корпуса) послужило основой для создания грамматик и словарей Cobuild. Было обнаружено, что корпусы позволяют проводить непосредственные наблюдения над текстами, включая анализ частотности слов и словосочетаний. При поиске устойчивых словосочетаний используются различные статистические методы, такие как критерий взаимной информации. Также, корпусы позволяют находить наиболее часто употребляемые словосочетания. Это важно для составления грамматических и лексических описаний языка. Однако, статистика взаимной информации отличается от простого упорядочивания словосочетаний по частотности. Она учитывает насколько совместное употребление двух слов отличается от случайного. Таким образом, корпусы позволяют обнаружить как редкие, так и часто встречаемые словосочетания. Введение статистических методов при работе с корпусами дает возможность более точно анализировать словосочетания и лексические единицы в языке. В дальнейшем, можно использовать более сложные статистические методы или комбинировать критерии статистики и взаимной информации, чтобы исключить незначимые или слишком часто встречаемые словосочетания. тексты на русском языке, возникает необходимость учитывать этот фактор при создании представительного корпуса. Необходимо определить временные рамки выборки, учитывая существенные изменения, произошедшие в языке на протяжении времени. Одной из важных задач при создании представительного корпуса является широкое покрытие различных типов текстов и функциональных стилей. Для этого была разработана типология текстов, основанная на рекомендациях синклера. Она основана на логических свойствах коммуникации и может быть адаптирована для русского дискурса. Важными внутренние факторы. Внешние факторы относятся к созданию текста автором, внешним признакам текста и целям создания. Внутренние факторы отражают предметную область текста и стилистические особенности. Также важно учитывать параметры создания текста, такие как время создания, возраст автора, пол и регион происхождения. Влияние региона на тексты может быть ограничено грубой классификацией (столичный, европейский, сибирский и южный), а возраст может быть разделен на детский, молодежный, взрослый и пожилой. Выбор временных рамок для корпуса также важен для отражения современного русского языка. В ходе исторического процесса в СССР и России язык изменился, поэтому выбор хронологических рамок существенно влияет на результаты исследования. Например, слова, связанные с советской эпохой, могут иметь высокую частоту в корпусе, но в современном использовании оказываются редкими.  
   факторами, влияющими на выбор текстов в корпусе, являются внешние и   
   Учет всех этих факторов и разработка представительного корпуса русского языка являются сложной задачей, но крайне важными для дальнейшего изучения и анализа русского языка. Описанные выше принципы построения представительных корпусов включают:1. Балансировку коллекции текстов по пять факторов классификации, что позволяет представить разнообразие языковых особенностей и жанров в корпусе. 2. Учет размера текста как важного технического параметра. Большой корпус может быть собран из длинных текстов, но для более полного представления различных языковых характеристик рекомендуется включение более коротких текстов. 3. Учет всех возможных комбинаций параметров для представления функциональных жанров русской речи. Количество текстов в каждой группе зависит от количества соответствующих текстов в русском дискурсе и их доступности в электронной форме. 4. Описание каждого текста в рамках базовых параметров, включая предметную область, стиль, авторство, тип публикации, аудиторию и коммуникативные цели. 5. Учет доступности текстов в электронном виде. Включение различных типов текстов, в том числе "эфемерных жанров", которые отражают современное использование русского языка, рассматривается как важная задача при создании представительного корпуса. Важно, чтобы каждый текст в корпусе соответствовал принципам балансировки и был описан в рамках основных параметров. При составлении корпуса учитывается представительность различных жанров и доступность текстов в электронной форме. Корпус стремится отражать норму современного русского литературного языка, поэтому в него могут быть включены в первую очередь художественная проза и тексты, написанные после 1960 года. Разметка текстов включает в себя лемматизацию и морфосинтаксическую разметку. Для русского языка, помимо разметки по частям речи, также необходимо проводить лемматизацию для облегчения поиска словоформ. Кроме того, грамматические характеристики слова (например, падеж, род, число) также помогут анализировать предложное и глагольное управление. Частичная синтаксическая разметка также предусмотрена в корпусе, поскольку автоматический анализ именных и предложных групп в русском языке достаточно надежен. Словоформы с неоднозначными грамматическими характеристиками составляют около 60-70% словоупотреблений, а словоформы с неоднозначным выделением лексемы и части речи - около 30-70%. Частичный синтаксический анализ может помочь в разрешении различных видов неоднозначности, например, в падежных формах, субстантивированных прилагательных, притяжательных и личных местоимений, а также в омонимии между существительными и глаголами в повелительной форме. Однако оставшуюся неоднозначность не удастся разрешить автоматически. Некоторые случаи омонимии и неоднозначности требуют полного синтаксического анализа на уровне предложения, а также семантического и прагматического анализа полного текста. Примеры неоднозначности включают омонимию между существительными, например, "поле" и "пол", а также между кратким прилагательным и наречием. Эти случаи требуют ручного разрешения. Размеченный эталонный корпус предполагает ручное разрешение оставшейся неоднозначности, что может быть длительным процессом. При принятии решений учитывается частота форм, например, спина почти всегда анализируется как существительное женского рода. Некоторые типы неоднозначности по выбору леммы и части речи остаются. Морфосинтаксическая разметка основана на формате тегов, где фрагменты текста помечаются с помощью тегов <s>, <w> и <апа>. Тег <апа> используется для присваивания различных видов анализа в поле анализа. Это позволяет представлять однозначные и омонимичные разборы одинаковым образом. Некоторая неоднозначность остается в разметке текстов и не может быть разрешена с помощью простых методов. Ручное разрешение неоднозначности требует времени и усилий, и некоторые решения основываются на частоте форм и других факторах. В заключение, можно сделать вывод, что корпусная лингвистика получила свое развитие преимущественно на материале английского языка. Это объясняется не только развитием компьютерной техники в США, но и интеллектуальным климатом в британской лингвистике, где были сильно выражены эмпирические тенденции. Результатом этого стало создание множества корпусов и проведение корпусных исследований на материале английского языка. Однако, эмпирический подход также присутствовал и в российской лингвистике, где лингвистический анализ всегда сопровождался примерами реального словоупотребления. Темы, поднимаемые исследователями в области корпусной лингвистики, часто созвучны с классическими темами российской лингвистики, связанными с исследованием употребления слов, выражений, грамматических конструкций и типов развития дискурса носителями языка в различных контекстах.
4. Сферы применения звуковых корпусов текста: 1. Лингвистика: - Фонетические исследования - Изучение произношения в различных языках - Исследование диалектов и акцентов 2. Речевые технологии: - Распознавание и синтез речи - Разработка голосовых помощников - Улучшение аудио-качества в приложениях 3. Психолингвистика: - Изучение взаимосвязи между языком и психологическими процессами - Анализ речевых нарушений и их влияние на психику. Практическая значимость звуковых корпусов текста: - Позволяют исследователям работать с большим объемом аудиоданных эффективно и систематически - Содействуют автоматизации процессов распознавания и анализа речи - Помогают в разработке технологий, например, в улучшении систем распознавания речи и создании голосовых ассистентов. Историческая значимость звуковых корпусов текста: - Звуковые корпуса текста помогают изучать и сохранять языковое наследие - Позволяют исследователям сравнивать и анализировать языковые структуры различных эпох - Содействуют сохранению и изучению уникальных диалектов и языковых особенностей