УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«ВИТЕБСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ П.М. МАШЕРОВА»

Факультет гуманитарного знания и коммуникаций

Кафедра белорусской и русской филологии

СОГЛАСОВАНО
Заведующий кафедрой
_____ Т.П. Слесарева
26.09.2024

СОГЛАСОВАНО
Декан факультета
_____ С.В. Николаенко
26.09.2024

УЧЕБНО-МЕТОДИЧЕСКИЙ КОМПЛЕКС
ПО УЧЕБНОЙ ДИСЦИПЛИНЕ

ИНФОРМАЦИОННАЯ ЛИНГВИСТИКА

для специальности

7-06-0232-01 Языкознание

Составитель: В.М. Генкин

Рассмотрено и утверждено
на заседании научно-методического совета 24.10.2024, протокол № 1

УДК 004.5:81(075.8)
ББК 81с51я73
И74

Печатается по решению научно-методического совета учреждения образования «Витебский государственный университет имени П.М. Машерова». Протокол № 1 от 24.10.2024.

Составитель: доцент кафедры белорусской и русской филологии ВГУ имени П.М. Машерова, кандидат филологических наук, доцент **В.М. Генкин**

Р е ц е н з е н т ы :
кафедра иностранных языков УО «ВГТУ»;
заведующий кафедрой германской филологии ВГУ имени П.М. Машерова, кандидат филологических наук, доцент *О.В. Шеверинова*

Информационная лингвистика для специальности 7-06-0232-01 Языкознание : учебно-методический комплекс по учебной дисциплине / сост. В.М. Генкин. – Витебск : ВГУ имени П.М. Машерова, 2024. – 71 с.

ISBN 978-985-30-0193-8.

В данном издании представлены сведения по одному из новых направлений современного языкознания – информационной лингвистике. Наряду с теоретическим блоком содержатся задания для практических занятий, контрольный тест, вопросы для подготовки по учебной дисциплине и другие материалы, призванные сделать образовательный процесс максимально эффективным. Большой список литературы позволит читателям расширить свои знания и получить дополнительную информацию по данной дисциплине и смежным с нею направлениям.

Адресованный в первую очередь магистрантам филологических специальностей, учебно-методический комплекс может быть использован преподавателями УВО, учителями гимназий и лицеев, а также всеми, кто интересуется современной лингвистикой.

УДК 004.5:81(075.8)
ББК 81с51я73

ISBN 978-985-30-0193-8

СОДЕРЖАНИЕ

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Коммуникация, осуществляемая с помощью новейших и постоянно развивающихся технологий, стала неотъемлемой частью информационной жизни каждого человека. Параллельно с этим повышается актуальность научного изучения новых форм коммуникации. Функционирование и использование языка в новых сферах обусловило появление ряда проблем теоретического и прикладного характера, связанных с разными аспектами познания, коммуникации, информационной обработки лингвистического материала. Решению этих проблем призвана способствовать инновационная дисциплина – информационная лингвистика, в сферу которой входит компьютерно-опосредованная коммуникация, интерпретация, репрезентация и моделирование информации. Очевидно, что цифровая грамотность сегодня является востребованной в сфере новейших компьютерных технологий коммуникации. Специалист должен уметь структурировать, обобщать, создавать и применять метаописания современной коммуникации, которые активно расширяются в сфере практической и научной деятельности.

Курс «Информационная лингвистика» входит в число дисциплин учреждения высшего образования.

Актуальность курса объясняется теоретической значимостью решения проблем коммуникативных аспектов лингвистики, а также потребностью дальнейшей разработки теории информационного анализа современного дискурса.

Цель дисциплины – углубить и расширить знания магистрантов об особенностях и перспективах современных коммуникативных процессов, дать полное и разностороннее представление об информационной лингвистике и ее месте в парадигме современных наук.

Курс «Информационная лингвистика» призван решить следующие **задачи:**

1) познакомить магистрантов-филологов с широким спектром коммуникативных исследований в языкознании, основными концепциями и фундаментальными научными трудами основоположников этой науки;

2) расширить знания о новых реалиях коммуникации;

3) сформировать понятие о лингвоинформационной специфике коммуникации;

4) рассмотреть особенности интернет-дискурса в аспекте моделирования коммуникации;

5) изучить институциональный, дискурсивный аспект функционирования коммуникации.

Теоретическую и языковую основу курса создают такие дисциплины, как «Современные направления языкознания», «Общее языкознание», «Лингвосемиотика», «Основы теории коммуникации», «Прагматика».

Освоение учебной дисциплины должно обеспечить формирование компетенции СК: применять компьютерные модели языка в лингвистике и смежных дисциплинах.

В ходе изучения курса магистрант должен:

знать:

– объект, предмет, задачи, основные разделы и методы исследования информационной лингвистики;

– базовые лингвистические понятия и термины информационной лингвистики (дискурс, коммуникация, информация, компьютерно-опосредованная коммуникация и т.д.);

– специфику лингвистических, лингвосемиотических и других исследований в области компьютерно-опосредованной коммуникации;

– сущность современных информационно-коммуникативных процессов;

уметь:

– свободно ориентироваться в различных направлениях информационной лингвистики;

– давать определение основным терминам информационной лингвистики и точно употреблять их в собственном выступлении на заданную тему;

– интерпретировать, моделировать информацию и осуществлять ее репрезентацию;

– осуществлять структурирование информации с учетом ее специфических особенностей;

– интегрировать знания из разных отраслей профессиональной деятельности;

– создавать и использовать метаописания современной коммуникации;

– использовать в лингвистических исследованиях компьютерно-информационные технологии;

– совершенствовать умения и навыки работы с лингвистической литературой;

иметь навык:

– лингвистического и дискурсивного анализа в рамках основных концепций данного направления при помощи различных методов и с использованием различных подходов, оперирования языковым материалом и примерами;

– использования базового терминологического инструментария современной информационной лингвистики;

– творческого использования знаний и их развития в ходе решения исследовательских и профессиональных задач.

ТЕОРЕТИЧЕСКИЙ РАЗДЕЛ

МОДУЛЬ I

Тема 1
ИНФОРМАЦИЯ И ЛИНГВИСТИКА

План:

1. Прикладная лингвистика и ее задачи.
2. Лингвистика и информация, лингвистика и информатика.
3. Методы прикладной лингвистики.

Ключевые понятия: *информация*, *информатика*, *прикладная лингвистика, теоретическая лингвистика*, *метод, моделирование*.

1. Прикладная лингвистика и ее задачи

В традиционном понимании лингвистикой, или языкознанием, принято называть «науку о естественном человеческом языке вообще и о всех языках мира как индивидуальных его представителях» [28, с. 618]. При этом естественный язык является наиболее важной и наиболее изученной знаковой системой. Лингвисты тщательно изучают строение языка (выделение в нем фонетического, лексического, грамматического уровня и уровня текста), социального варьирования языка, вопросы:

– порождения и понимания языковых высказываний;

– принципы функционирования языка в обществах разных типов;

– происхождения и развития языка.

Лингвистику, естественно, интересуют другие аспекты строения и функционирования языка. В зависимости от изучаемого аспекта, национальной традиции и научной методологии выделяются различные разделы лингвистики, в числе которых структурная лингвистика, социолингвистика, психолингвистика и т. п.

В языкознании, как и в других науках, принято разграничивать теоретическое и прикладные направления. Теоретическая лингвистика, которую еще называют фундаментальной, представляет собой науку, изучающую объективное состояние языка, его историю, сложившиеся в нем закономерности и т.д. В широком смысле она должна дать ответ на вопрос «Каков язык?»

Прикладная лингвистика – это «направление в языкознании, занимающееся разработкой методов решения практических задач, связанных с использованием языка» [28, с. 397]. Вопрос, на которой она должна ответить,

звучит так: «Как лучше использовать язык?» Следует отметить, что в зарубежном и восточнославянском языкознании термин «прикладная лингвистика» употребляется в разных значениях. В 1930–40-е гг. английское название ʻapplied linguisticsʼ стали употреблять применительно к процессу обучения иностранному языку, включая в него методику преподавания иностранного языка, особенности описания грамматики в учебных целях и т.п.

В русскоязычной традиции начиная с 1950-ых гг. под прикладной лингвистикой понимали компьютерные технологии, автоматические системы обработки информации и т.д. Здесь синонимами понятия ʻприкладнаяʼ будут выступать прилагательные ʻкомпьютернаяʼ, ʻвычислительнаяʼ, ʻавтоматическаяʼ, ʻинженернаяʼ. Другими словами, информационная лингвистика и есть собственно прикладная лингвистика в русскоязычном терминологическом употреблении.

К числу традиционных задач прикладной лингвистики, понимаемой в широком смысле, относятся:

– создание и совершенствование письменностей;

– создание систем транскрипции устной речи;

– создание систем транслитерации иноязычных слов;

– создание систем стенографии;

– создание систем письма для слепых;

– упорядочение, унификация и стандартизация научно-технической терминологии;

– изучение процессов и создание правил образования названий новых изделий, товаров, химических веществ;

– разработка методов адекватного преобразования текстов в иноязычную форму (перевода);

– совершенствование методики преподавания языков и др.

Развитие компьютерных технологий поставило перед прикладной (информационной) лингвистикой новые задачи, в числе которых:

– разработка лингвистических основ машинного перевода;

– автоматическое индексирование и аннотирование документов;

– автоматический анализ текстов;

– автоматический синтез текстов;

– создание словарей-тезаурусов для автоматического поиска информации и др.

Современная прикладная лингвистика включает в себя такие направления, как *лингвистические основы информатики* и *компьютерная лингвистика*. Теоретическую базу прикладной лингвистики создает лингвистика фундаментальная. По ряду позиций эта база еще недостаточно разработана. Так, наиболее проблемными теоретическими аспектами являются вопросы владения языком, «препарирования» языка для кибернетических целей и некоторые другие. Основываясь на многовековом опыте изучения языка, прикладная лингвистика активно и успешно использует развивающиеся возможности современной техники и технологий.

2. Лингвистика и информация, лингвистика и информатика

Современный мир активно сменяет статус общества индустриального на общество информационное. Все большую и бóльшую роль в нем играет информация, которая, согласно известному выражению, правит этим миром. «Основным перерабатываемым «сырьем» становится информация. Труд современников делается в меньшей степени физическим и в большей степени интеллектуальным. В наиболее развитых странах производство информации и разработка информационных технологий стало одной из самых прибыльных и стремительно растущих отраслей» [12, с. 7].

Слово ‘информация’, восходящее к латинскому существительному *informatio – разъяснение, изложение*, сегодня используется в двух основных значениях:

1. «Сообщение о фактах, событиях, о состоянии чего-либо.

2. Совокупность сведений как объект хранения, переработки и передачи» [40, с. 160].

Во втором значении информация стала естественным объектом изучения ряда наук, в частности когнитивных, а также теории коммуникации, компьютерной лингвистики, информатики, кибернетики, теории проектирования и др.

Стратегической целью непосредственно компьютерной лингвистики является решение задач лингвистического обеспечения информатики, которую обычно понимают как «науку о закономерностях записи хранения, переработки, передачи информации с помощью современных технических средств. Поскольку мы живем в компьютерный век, имеется в виду осуществление этих процессов в основном с помощью компьютера» [34, с. 13].

Термин ‘информатика’ происходит от французского слова *informatique*, употребляемого и в роли прилагательного (*компьютерный, машинный, информационный*), и в роли существительного (*вычислительная техника, компьютерная техника, информатика*). Современные толковые словари дают следующие дефиниции этого понятия:

«1. Отрасль науки, изучающая общие свойства информации, а также вопросы, связанные с её накоплением, преобразованием, поиском, хранением и передачей с помощью компьютеров и других вычислительных средств.

2. Сфера практического применения вычислительной техники» [40, с. 160].

К области информатики относят:

software – программное обеспечение;

hardware – аппаратное обеспечение;

lingware – лингвистические обеспечение.

Именно последний компонент является зоной прямого пересечения информатики и лингвистики, помогающей создавать специальные словари, алгоритмы анализа текстов и способы их автоматического создания.

Без компьютерной лингвистики и информатики невозможна современная массовая коммуникация – «общезначимый современный текст, в создании и распространении которого принимают участие новейшие технические средства и устройства: мощные печатные машины, телевидение, кино, магнитофонная запись, компьютеры и пр. Причем это преимущественно текст серьезного характера, служащий главным образом нуждам общественного управления, связанный с развитием, регулированием и устройством современного массового производства» [53, с. 163].

Для обозначения технических средств используется терминологическое сочетание ʻноситель информацииʼ – «материальный объект, предназначенный для хранения данных» [40, с. 160].

Еще одним важным аспектом изучения и использования информации стала информационная культура (education culture) – «умение целенаправленно работать с информацией и использовать ее для получения, обработки и передачи компьютерную информационную технологию, современные технические средства и методы» [12, с. 8].

Информационная культура проявляет себя в:

1) конкретных навыках по использованию технических устройств (от телефона до персонального компьютера и компьютерных сетей);

2) способности использовать в своей деятельности компьютерную информационную технологию, базовой составляющей которой являются многочисленные программные продукты;

3) умении извлекать информацию из различных источников: как из периодической печати, так и из электронных коммуникаций, представлять ее в понятном виде и уметь ее эффективно использовать;

4) владении основами аналитической переработки информации; в умении работать с различной информацией; в знании особенностей информационных потоков в своей области деятельности [12, с. 8].

3. Методы прикладной лингвистики

Методы прикладной лингвистики очень разнообразны, что в первую очередь обусловлено многообразием областей приложения лингвистического знания. Нужно отметить, что каждая конкретная прикладная дисциплина имеет свой набор методологических инструментов. Так, например, «квантитативная лингвистика в значительной мере опирается на методический инструментарий статистики, компьютерная лингвистика широко использует методы теории программирования и представления знаний, теория воздействия опирается на представление о значимом варьировании языковых структур» [5, с. 6]. Квантитативная лингвистика представляет собой направление, которое исследует язык с помощью точных математических методов и компьютерных программ.

Описательная лингвистика ориентирована на классификационное представление языковых явлений, то есть «выявление той сетки параметров,

которая позволяет охватить все релевантные (в теории) свойства языковых структур» [5, с. 7]. Метод классификации находит применение и в структурных исследованиях, где он, естественно, имеет свою специфику.

Теоретическое языкознание представляет язык в его концептуальных моделях, которые должны не только описать и расклассифицировать лингвистические факты, но и дать им должное толкование. Описательная лингвистика, помимо прочего, может использовать и *метод моделирования*, который эффективно использовался в структурной лингвистике, потому что именно моделирование структурных явлений приводит исследователя к требуемому результату. Относительно явлений, не обладающих ярко выраженной структурной организацией, применение этого метода сопряжено с таким затруднением, как невозможность прямого перенесения модели на сам изучаемый объект.

Наиболее очевидное условие выбора метода моделирования – наличие препятствий для непосредственного изучения того или иного объекта исследования. В качестве примера подобного объекта можно привести мышление человека и связи мышления с языком. «Исходный пункт использования метода моделирования – представление о входной и выходной информации (в широком понимании), характеризующей функционирование объекта моделирования» [5, с. 7]. Применительно к естественному языку метод моделирования может быть продуктивным, например, для изучения синтаксиса.

Верификация созданной модели осуществляется следующим образом: ее сравнивают с поведением объекта моделирования. При условии их совпадения, наличия закономерной повторяемости созданная модель считается успешной, «теория моделирования позволяет в этом случае перенести особенности устройства модели на сам объект» [5, с. 7].

Исследователи говорят о наличии ограничений при использовании метода моделирования. Ю. Д. Апресян в вышедшей еще в 1966 году работе «Идеи и методы современной структурной лингвистики» подчеркивал, что моделировать можно только те свойства объекта, которые не определяются его физической природой» [4, с. 79]. Ученый характеризует модель как идеализацию объекта, которая должна предсказывать поведение объекта и объяснять его [4, с. 79].

Теоретическая лингвистика, используя метод моделирования, ориентируется на несколько типов моделей, в числе которых модели:

– *компонентные* (модели структуры), представляющие, соответственно, внутреннюю структуру объекта;

– *предсказывающие*, способные дать информацию прогнозирующего характера, т.е. предсказать поведение объекта в тех или иных обстоятельствах;

– *имитирующие*, демонстрирующие особенности внешнего поведения объекта;

– *диахронические*, призванные продемонстрировать исторические изменения в структуре объекта и объяснить их.

Метод моделирования еще более эффективен в прикладных исследованиях, задачей которых является оптимизация определенных лингвистических функций. Оптимизация предполагает создание таких моделей проблемной области, которые способны представить ее существенные свойства, необходимые для решения поставленных задач. В отличие от теоретической лингвистики, стремящейся к полному описанию языковых явлений и фактов, прикладная – ставит целью их оптимизированное описание, которое «должно быть удовлетворительным для данной конкретной задачи» [5, с. 7].

В прикладной лингвистике применяется так называемый инженерный подход, ориентированный на познавательную установку, как раз и связанной с решением конкретной задачи. «Это, впрочем, не означает, что результаты прикладных исследований не представляют никакой ценности для теории языка: напротив, прикладные модели оказывают значительное влияние на лингвистическую теорию, способствуя обновлению концептуального аппарата современного языкознания» [5, с. 7].

Тема 2
ЛИНГВИСТИЧЕСКИЕ ОСНОВЫ ИНФОРМАТИКИ И КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

План:

1. Моделирование общения.
2. Метауровень коммуникации и его роль в моделировании общения.
3. Типы коммуникативного общения и их учет в моделировании коммуникации.
4. Моделирование сюжета.
5. Синтаксис сюжета и когнитивный подход к моделированию.

Ключевые понятия: *коммуникация, уровни коммуникации, метакоммуникация, моделирование, сюжет*.

1. Моделирование общения

Как уже отмечалось, одним из наиболее продуктивных методов современной прикладной лингвистики является *метод моделирования*. Моделированием называют «исследование каких-либо объектов на их моделях; построение моделей реально существующих предметов, явлений или процессов (живых организмов, инженерных конструкций, образцов одежды и т.п.)» [40, с. 243].

Одной из задач компьютерной лингвистики является *моделирование общения*, в частности общения человека и ЭВМ, в основе которого лежит естественный или ограниченный естественный человеческий язык. При этом компьютерные модели общения используются для изучения самого процесса общения. Исследователи отмечают, что накопленный опыт

эксплуатации компьютерных систем позволил по-новому взглянуть на естественную человеческую коммуникацию, на ее функции и структуру.

Сегодня модели общения человека с компьютером изучаются в рамках специального направления науки об искусственном интеллекте, которое имеет англоязычное название *Human-Computer Interaction* (HCI). В числе исследовательских задач этого направления находится моделирование дискурса и дискурсивных стратегий. В центре внимания оказываются вопросы и проблемы, которые ранее были на периферии теорий диалога, дискурс-анализа и теории коммуникации. Современные исследователи в области прикладной лингвистики пытаются найти ответы на следующие вопросы:

– Что обеспечивает естественность общения?

– Каковы условия связности беседы?

– Когда общение оказывается успешным?

– В каких случаях возникают коммуникативные неудачи и можно ли их избежать?

– Какие стратегии общения используют участники диалога для достижения своих коммуникативных целей?

В 1966 году был создан первый вариант программы Джозефа Вейценбаума «Элиза», которая представляла собой компьютерную модель диалога (Описание программы и проведенного на ее основе эксперимента приводится по учебнику А. Н. Баранова «Введение в прикладную лингвистику» [5]).

Имя для программы было выбрано Дж. Вейценбаумом не случайно: в пьесе Б. Шоу «Пигмалион» профессор Хиггинс учит Элизу Дулитл говорить на литературном английском языке. Изначально программа задумывалась как игрушка, как учебный имитатор речевого поведения. При этом создатель не ставил перед собой задачу моделирования мышления. Программа поддерживала разговор в режиме реального времени, но при ее разработке были использованы ограниченные программистские ресурсы, лингвистический анализ и синтез тоже были минимальны.

Программа «Элиза» была использована группой исследователей во главе с М. Макгайром для изучения структуры диалога и особенностей естественно-языковой коммуникации. Был поставлен эксперимент, в ходе которого 24 его участника в течение часа беседовали с «Элизой» с помощью телетайпа. Телетайп (англ. *teletype*) – «приемопередающий буквопечатающий телеграфный аппарат; применяется для передачи сообщений по каналам связи, а также в качестве терминала в устройствах вычислительной техники» [40, с. 402]. Терминалом (от англ. *terminal* < лат. *terminalis* – конечный, относящийся к концу) называется «устройство для ввода и вывода информации (клавиатура, дисплей, принтер и др.)» [40, с. 405].

Каждый участник ввел от 10 до 65 реплик и получил на них ответы. По окончании 15 участников (62%) были уверены, что общались с человеком, 5 (21%) участников отметили определенные проблемы и только 4 человека (17%) заявили уверенно, что общались с машиной. При этом алгоритмы

программы включали минимум лингвистической информации: во-первых, это комплекс ключевых слов, актуализирующих некоторые устойчивые формулы (шаблоны); во-вторых, способность относительно несложно трансформировать предшествующее высказывание (программа, например, могла переделать предшествующее выказывание в общий вопрос)

Одним из показателей успешности программы стало, по сути, опровержение известного теста А. Тьюринга на создание искусственного интеллекта. А. Тьюринг в свое время вместо обсуждения сложного философского вопроса о возможности машины мыслить предложил игровую задачу, условия которой были следующими: есть три участника (мужчина *Л*, женщина *В*, спрашивающий *С*). При этом спрашивающий не знает, кто мужчина, а кто женщина. Задавая вопросы, он должен определить пол собеседников, но участник *Л* пытается мистифицировать спрашивающего, предлагая ему не ложную, но искаженную информацию, а участник *В*, напротив, стремится помочь. Общение происходит не напрямую, а через телетайп или с помощью записок, отпечатанных на пишущей машинке. Возникает дополнительный вопрос: что произойдет, если в качестве участника *А* будет выступать система искусственного интеллекта?

Более простой вариант этой задачи сводится к тому, что несколько участников беседуют с некоторым другим участников *Х*. Проблема построения искусственного интеллекта будет считаться решенной, если большинство участников не смогут установить, с кем они беседуют: с человеком или машиной.

В одном из вариантов программы «Элиза» перевоплощается в доктора-психиатра (фрагмент диалога с программой «Доктор»):

«Доктор. *Говорите*!

Пациент. *Меня волнует погода*.

Доктор. *Как давно вас волнует погода?*

Пациент. *С тех пор, как умерла моя мать*.

Доктор. *Расскажите больше о вашей семье*.

Пациент. *Они все в области компьютеров*.

Доктор. *Машины пугают вас*.

Пациент. *Нет*.

Доктор. *Пожалуйста, не будьте столь кратки*.

Пациент. *Извините*.

Доктор. *Вы говорили о вашей матери*».

По внешним характеристикам перед нами предстает совершенно нормальная беседа врача с пациентом. Между тем даже поверхностный анализ показывает, что программа «Доктор» не понимает собеседника в точном смысле: она не имеет своей модели мира, не строит в ней отображение представлений/знаний собеседника, не модифицирует модель мира с каждым новым высказыванием, не накапливает информацию, хотя у нее есть некоторые аналоги кратковременной и долговременной памяти.

Так, программа выделяет слово ‘мать’ и способна вставить его в дальнейшем в речевой шаблон. *Мать* для машины – это некий упомянутый *X,* который она использует в шаблоне: «Вы говорили об *X*» (так называемая косвенно-инициирующая реплика, или косвенный вопрос). В программу заложены и другие шаблоны: «Расскажите больше об *X*»; «Как давно вас волнует …?»; «Пожалуйста, не будьте столь кратки». Эти шаблоны построены таким образом, чтобы побудить адресата продолжить общение. Показательно, что существенная тематическая ограниченность коммуникации и значительное количество ошибок и неточностей в ответе (исследователи насчитали примерно 19% неточных или выпадающих из контекста реплик) не помешали испытуемым признать партнера по коммуникации человеком. Во многом эта ситуация объясняется терпимостью естественной коммуникации к сбоям и ошибкам. Испытуемые легко объясняли несовпадения обычными сбоями в понимании предшествующей реплики, не вполне нормальными условиями общения, шутливым настроением партнера. Устойчивость естественного дискурса объясняется также способностями человека к интерпретации речевых действий. Имея установку на коммуникацию, человек, участвующий в диалоге, ведет себя соответствующим образом. Он стремится включать в коммуникацию все то, что по форме напоминает речевой акт, реплику. Можно сказать, что он склонен наделять смыслом даже то, что часто смысла не имеет. Другими словами, участник диалога способен породить некий смысл диалога, обеспечить его связность, приписать партнеру коммуникативные интенции.

За почти шестьдесят лет, прошедших с момента создания программы «Элиза», технологии изменились коренным образом. Современные модели обладают множеством способностей, которые, естественно, не могли присутствовать у программы той эпохи.

2. Метауровень коммуникации и его роль в моделировании общения

Из вышеописанного эксперимента напрашивается и второй важный вывод: испытуемые довольно быстро принимали решение о том, кто перед ними: человек или компьютер. Так, 22 человека из 24 определились с эти в пределах первых пяти реплик и далее не меняли своей точки зрения. Иными словами, они на самом начальном этапе общения устанавливали коммуникативные роли. Определение ролей в коммуникации относится к метауровню общения, так как составляет одну из предпосылок успешной коммуникации.

Понятия ‘метауровень’, ‘метакоммуникация’ содержат в себе греческую приставку *meta*, синонимом которой является русский префикс *после*. Так, например, метаданными в науке принято называть «данные, являющиеся описанием других данных (напр., схемы базы данных по соотношению к содержимому базы данных» [40, с. 237]. Метакоммуникацию часто называют коммуникацией по поводу коммуникации. Основу данного понятия составляет идея о том, что структура человеческой коммуникации является сложной.

Коммуникация осуществляется на разных уровнях, первый из которых относится к области информации, т.е. содержания общения, а последний классифицирует первый и представляет собой уровень метакоммуникации.

Известный английский исследователь Г. Бейтсон (1904–1980), рассматривая в работе «Экология разума» все человеческое поведение как коммуникацию, приходит к выводу, что при таком подходе, «мы будем иметь дело не с изолированной единицей сообщения, а с изменчивым и многогранным объединением многих форм поведения – вербальных, тональных, касающихся поз, контекстуальных и т.д., каждая из которых определяет смысл всех остальных» [6, с. 10].

Область простой коммуникации, т.е. непосредственной передачи информации, дополняется сопутствующими знаками-маркерами, которые по-своему расширяют информацию, кодируют или классифицируют ее, что создает контекст для коннотаций. Исследователи приводят примеры несложных метакоммуникативных сообщений, таких как «Это шутка», «Это игра», «это все по-настоящему», «Это угроза», «Я не шучу», «Это не угроза, а предупреждение» и т. п. Совершенно очевидно, что каждый из маркеров может значительно, если не кардинально, изменить характер передаваемой информации.

А. Н. Баранов, комментируя понятие метауровня коммуникации, иллюстрирует его ссылкой на фрагмент из романа И. Ильфа и Е. Петрова «Золотой теленок», где обращение к метауровню прекращает нормальный диалог и переводит в брутальную область.

«У Балаганова сразу сделалось мокрое, как бы сварившееся на солнце, лицо.

– Зачем же мы работали? – сказал он, отдуваясь. – Так нельзя. Это... объясните.

– Вам, – вежливо сказал Остап, – любимому сыну лейтенанта, я могу повторить только то, что я говорил в Арбатове. Я чту Уголовный кодекс. Я не налетчик, а идейный борец за денежные знаки. <…>

– Зачем же вы послали нас? – спросил Балаганов, остывая. – Мы старались...

– Иными словами, вы хотите спросить, известно ли достопочтенному командору, с какой целью он предпринял последнюю операцию? На это отвечу – да, известно. Дело в том...

В эту минуту в углу потух золотой зуб. Паниковский развернулся, опустил голову и с криком: «А ты кто такой?» – вне себя бросился на Остапа. Не переменяя позы и даже не повернув головы, великий комбинатор толчком каучукового кулака вернул взбесившегося нарушителя конвенции на прежнее место».

Реплика Паниковского *А ты кто такой?* в данном контексте является вовсе не требованием информации, а маркером перехода на метауровень общения – она связана с выяснением роли, статуса Остапа Бендера

в микросоциуме (образующем коммуникативную группу) Остапа и его коллег-подельщиков. Разумеется, переход на метауровень общения не обязательно связан с физическим конфликтом. Реплики представления (самопредставления) типа *Разрешите представиться,* приветствия и прощания также относятся к метакоммуникации. Многие институциональные процедуры типа заседания суда, защиты диссертации включают значительный метакоммуникативный компонент, выполнение которого формально необходимо для успешности процедуры. Так, проведение судебного заседания предполагает обязательное выяснение того, является ли ответчик «надлежащим» ответчиком, то есть тем лицом, которому действительно можно предъявлять какие-то претензии.

Понятно, что определение ролей участников во многом определяет выбор стратегии коммуникативного поведения. Действительно, лучше сразу определить, с кем мы разговариваем по телефону – с давним другом или чиновником налоговой инспекции. Выяснение того, кем является собеседник – машиной или человеком, также относится к метауровню общения, и испытуемые старались установить ролевые характеристики партнера как можно раньше.

Это свойство естественно-языковой коммуникации можно назвать *принципом приоритета метакоммуникативных параметров* ситуации общения [5, с. 14].

Из сказанного очевидно, что моделирование общение невозможно только на уровне передачи/приема информации. Коммуникация будет эффективной только в том случае, если полностью будет учтен метакоммуникативный уровень.

3. Типы коммуникативного общения и их учет в моделировании коммуникации

Важное следствие из эксперимента М. Макгайра связано с существованием различных типов коммуникативного взаимодействия между людьми. Успешное взаимодействие между человеком и программой типа «Элиза» возможно только в ситуации, когда происходит так называемое «ассоциативное общение», при котором реплики диалога связаны не столько логическими отношениями типа «причина–следствие», «посылка–заключение», а ассоциациями. Ассоциативное общение не имеет конкретной направленности; само поддержание беседы может служить ее оправданием. Собеседники не преследуют цели решить какую-то проблему или выработать единую точку зрения на какой-то вопрос. В классификации Р. Якобсона для коммуникации такого типа предложен термин «фатическое общение» [57, с. 47]. Заметим, что беседа врача-психиатра с пациентом по форме также имеет вид фатического общения, хотя и преследует вполне определенные цели сбора данных о заболевании пациента и последующем вербальном и невербальном воздействии на его психику для достижения лечебного эффекта.

«Элиза» не смогла бы успешно имитировать общение в коммуникативной ситуации, названной М. Макгайром «решение задач», поскольку она не способна понять проблемную ситуацию, то есть построить модель мира дискурса, определить альтернативы выхода из проблемы, выбрать одну из альтернатив и т.д.

Одна из типичных стратегий «ухода от непонимания», реализованная в программе «Элиза», – смена темы беседы. Очевидно, что такая стратегия ведения беседы вряд ли приведет к успеху при совместном поиске решения проблемы.

Наконец, четвертый вывод можно сформулировать как неуниверсальность правил коммуникативного взаимодействия. Это касается самих закономерностей общения на естественном языке.

Каждый тип коммуникации обслуживается своим набором относительно простых правил, обеспечивающих связность дискурса, его осмысленность для участников. Типология видов общения задается соответствующими наборами правил. Из экспериментов М. Макгайра с программой «Элиза» следует, что кроме ассоциативного (= фатического) способа общения, выделяется еще «решение задач», «задавание вопросов» и «уточнение понимания». С лингвистической точки зрения эти типы, скорее всего, неоднородны, пересекаются и даже находятся на разных уровнях дискурса. Так, «уточнение понимания» относится к метауровню коммуникации, «задавание вопросов» может быть частью стратегии «решение задач» и «уточнения понимания» и т.д. Существенно, что компьютерный эксперимент с программой, моделирующей поведение участника коммуникации, позволяет экспериментально подтвердить или опровергнуть многие положения теории диалога, разработанные как в лингвистике, так и в смежных дисциплинах – в дискурс-анализе, теории коммуникации, психологии и социологии общения [5, с. 13 – 14].

4. Моделирование структуры сюжета

Изучение структуры сюжета относится к проблематике структурного литературоведения, психологии творчества и культурологии. Сюжетом в традиционном смысле называют последовательность событий, происходящих в каком-либо художественном произведении. Моделирование сюжета предполагает обращение к формальной теории сюжета, созданной структуралистами.

Структурализм в литературоведении представляет собой направление и метод исследования, заимствованный у языкознания. Его задача – обнаружить, описать и объяснить структуры мышления, которые лежат в основе культуры прошлого и настоящего. Начало использования этого метода связывают с именем французского этнолога К. Леви-Строса. Позднее некоторые принципы и идеи литературного структурализма были использованы

для создания соответствующих компьютерных программ. Следует выделить три так называемых базовых *формализма* представления сюжета. Это его морфологическое представление, синтаксическое представление и когнитивный подход к сюжету.

Идеи о морфологическом устройстве структуры сюжета восходят к известным работам В.Я. Проппа о русской волшебной сказке [52]. Самой известной работой В.Я. Проппа (1895–1970) является исследование под названием «Морфология сказки» (Ленинград, 1928 г.), где рассматривается структура народных сказок. Ученый выделил в сказках постоянные и переменные элементы. Постоянными являются действия, совершаемые персонажами для развития сюжета, и их последовательность. К переменным относятся языковой стиль, количество действий, способы их исполнения, а также мотивировки и атрибуты персонажей.

В.Я. Пропп заметил, что при обилии персонажей и событий волшебной сказки количество функций персонажей ограничено: «Постоянными, устойчивыми элементами сказки служат функции действующих лиц, независимо от того, кем и как они выполняются. Они образуют основные составные части сказки» [52, с. 31]. К числу базовых относятся, например, следующие функции:

– отлучение персонажа сказки из дома;
– запрет герою на действие;
– нарушение запрета;
– получение вредителем информации о жертве;
– обман жертвы вредителем;
– невольное пособничество жертвы вредителю и т.д.

Идеи В.Я. Проппа легли в основу компьютерной программы TALE, моделирующей порождение сюжета сказки. В основу алгоритма программы TALE положена последовательность функций персонажей сказки. Фактически функции, описанные В.Я. Проппом, задавали множество типизированных ситуаций, упорядоченных на основе анализа эмпирического материала. Возможности сцепления различных ситуаций в правилах порождения определялись типичной последовательностью функций – в том виде, в котором это удается установить из текстов сказок. В программе типичные последовательности функций описывались как типовые сценарии встреч персонажей.

В дальнейшем система была усложнена за счет введения модели мира сказки, география которого состоит из обычного мира, промежуточного (среднего) мира и иного мира.

Каждый мир состоит из локусов, связанных между собой определенными отношениями. Отношения связывают не только локусы внутри каждого мира, но и локусы различных миров. Обычный мир состоит из следующих локусов: место проживания героя (локус 1), место получения задания (локус Г), место дарения волшебных предметов, помогающих выполнить задание.

Первый локус и локус штрих часто совпадают (ср. сказки о Падчерице и злой Мачехе). К обычному миру относятся также локусы 3 (их может быть много), в которых преодолеваются препятствия с помощью волшебных предметов. Количество препятствий, как правило, совпадает с количеством волшебных предметов. После преодоления препятствий герой оказывается в промежуточном мире, стражем которого является Баба-Яга. Средний мир отделяет мир героев от мира антигероев. Функции Бабы-Яги различаются – она может выступать как дарительница информации или очередного волшебного средства, а может выступать на стороне антигероев (например, при акценте на людоедском поведении Бабы-Яги). Иной мир включает место обитания антигероя (локус 5), место битвы между героем и антигероем (локус 6) и, наконец, локус 7 – место награды или цели, которой добивается герой. Локусы связаны отношениями перехода, которые представляют возможные последовательности развертывания сюжета. Модифицированная версия программы TALE имеет следующую блок-схему.

Работа программы начинается с первого блока, в котором выбирается тип сюжета сказки и ее персонажи. Здесь же формируется экспозиция сказки (setting). Во втором блоке хранятся описания, связанные с персонажами, а в четвертом – постоянные характеристики персонажей. Описания даются во фреймоподобных структурах представления знаний. С помощью второго и третьего блоков формируются мотивы и поступки персонажей. Третий блок задает последовательность движения персонажей по локусам.

В последнем (шестом) блоке происходит сборка порожденных фрагментов сказки.

Блок-схема модифицированного варианта программы TALE показывает, что чисто «морфологического» подхода к структуре сюжета сказки явно недостаточно. «Морфемы» сказочного сюжета должны не только определенным образом сочетаться между собой, но и иметь специфические ограничения на сочетаемость. Фиксация одного типичного порядка следования функций персонажей волшебной сказки существенно ограничивает имеющиеся возможности сочетаемости. Более адекватное решение этой проблемы дает синтаксический подход к структуре сюжета.

5. Синтаксис сюжета и когнитивный подход к моделированию

Теоретическую основу синтаксического подхода к сюжету текста составили «сюжетные грамматики» (story grammars). Сюжетные грамматики появились в середине 70-х гг. в результате переноса идей порождающей грамматики Н. Хомского на описание макроструктуры текста. Если важнейшими составляющими синтаксической структуры в порождающей грамматике были глагольные и именные группы, то в большинстве сюжетных грамматик в качестве базовых выделялись экспозиция (setting), событие

и эпизод. В теории сюжетных грамматик широко обсуждались условия минимальности: ограничения, определявшие статус последовательности из элементов сюжета как нормальный сюжет.

Оказалось, однако, что чисто лингвистическими методами это сделать невозможно. Многие ограничения носят социокультурный характер. Сюжетные грамматики, существенно различаясь набором категорий в дереве порождения, допускали весьма ограниченный набор правил модификации нарративной структуры. В подавляющем большинстве случаев эти правила заимствованы из той же порождающей грамматики.

Нарратив (англ., фр. *narrative* – лат. *narrare* ʻрассказывать, повествовать’) – это самостоятельно созданное повествование о некотором множестве взаимосвязанных событий, представленное читателю или слушателю в виде последовательности слов или образов. Часть значений термина «нарратив» совпадает с общеупотребительными словами «повествование», «рассказ». Сегодня существует даже специальное учение о нарративе – нарратология. С начала XXI века под влиянием англоязычной политологии термин приобрёл также и в русском языке дополнительное значение – «высказывание, которое содержит мировоззренческую установку/предписание». Это значение синонимично понятию ʻидеологема’

Потенциал варьирования структуры сюжета обеспечивается в первую очередь трансформациями передвижения и опущения. Например, текст признания преступника, фиксирующий реальную последовательность развертывания событий в преступлении, можно с помощью перестановок и опущений преобразовать в детективный сюжет:

{преступник → замысел → орудие убийства → место → убийство → обнаружение трупа → поиски преступника} → {обнаружение трупа → обнаружение орудия убийства → поиски преступника}.

Использование сюжетных грамматик в компьютерном моделировании оказалось не вполне удачным. Синтактический компонент сюжета, описываемый грамматиками, отражает чисто внешние особенности текста. Не удается обнаружить операциональные критерии выделения различных составляющих сюжета.

Например, где в сюжете эпизод, а где событие? Попытка использовать грамматики сюжетов для порождения сюжета приводит к тому, что порождаются тексты, которые не отвечают интуитивному представлению о рассказе. Например, терминальная цепочка, порождаемая одной из грамматик обсуждаемого типа, «Экспозиция + Тема + Сюжет + Разрешение» вполне может быть приписана предписаниям-советам следующего вида: «Вас позвали на рыбалку, а вы ничего в этом не смыслите. Что ж, сначала вам надо обзавестись рыболовными снастями. Вы можете пойти в магазин и купить спиннинг. Чтобы выбрать хороший спиннинг, надо...» [Black, Bower 1980].

Основной вывод дискуссии о недостатках сюжетных грамматик свелся к необходимости описания сюжета в рамках структуры целесообразной деятельности, то есть с привлечением категорий «цель», «проблема», «план» и т.д. Иными словами, метаязыка, учитывающего только внешние особенности сюжета, явно недостаточно. Необходимо обращение к когнитивным состояниям персонажей.

Когнитивный подход разрабатывался разными исследователями, среди которых особо следует выделить американского теоретика искусственного интеллекта Роджера Карла Шенка (1946 – 2023). Одна из учениц Р. Шенка Венди Грейс Ленерт, ученый-компьютерщик, специализирующийся на обработке естественного языка, стала известной благодаря своему новаторскому подходу в использовании машинного обучения в обработке естественного языка.

В начале 80-х гг. В. Ленерт в рамках работ по созданию компьютерного генератора сюжетов был предложен оригинальный формализм аффективных сюжетных единиц (ACE – Affective Plot Units), оказавшийся мощным средством представления структуры сюжета. При том что он был изначально разработан для системы ИИ, этот формализм использовался в чисто теоретических исследованиях. Сущность подхода В. Ленерт заключалась в том, что сюжет описывался как последовательная смена когнитивно-эмоциональных (аффективных) состояний персонажей. Тем самым в центре внимания формализма В. Ленерт стоят не внешние компоненты сюжета – экспозиция, событие, эпизод, мораль – а его содержательные характеристики. В этом отношении формализм Ленерт отчасти оказывается возвращением к идеям Проппа.

Каждая аффективная сюжетная единица представляет собой бинарное отношение, связывающее некоторые события, оцениваемые персонажами положительно (+) или отрицательно (–), и когнитивно-эмоциональные состояния персонажей (в различных комбинациях – событие & состояние; событие & событие и т.д.). Бинарное отношение не однородно. Всего выделяется пять типов бинарных отношений, специфицируемых в каждой аффективной сюжетной единице. Бинарное отношение может быть мотивацией (обозначение – т), актуализацией (a), прекращением одного действия другим (t), эквивалентностью (e), а также аффективной каузальной связью между персонажами. Каждая аффективная сюжетная единица получает название, например, УСПЕХ, НЕУДАЧА, УПОРСТВО, ПРОБЛЕМА и т.д. В разных вариантах формализма выделяется от 20 до 60 простых и комплексных аффективных сюжетных единиц.

Тема 3
КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

План:

1. Формирование компьютерной лингвистики как научного направления.
2. Межпредметные связи компьютерной лингвистики.
3. Теоретическое и прикладное направление компьютерной лингвистики.

Ключевые понятия: *компьютерная лингвистика*, *информация*, *интеллект*, *искусственный интеллект*, *прикладная лингвистика, теоретическая лингвистика*.

1. Формирование компьютерной лингвистики как научного направления

Компьютерная лингвистика представляет собой одно из наиболее активно разрабатываемых направлений современного языкознания, которое затрагивает области математического и компьютерного моделирования интеллектуальных процессов у человека и животных при создании систем искусственного интеллект и ставит целью использование математических моделей для описания естественных языков.

Прилагательное ‘компьютерная’ в составе названия дисциплины восходит к английскому *computer* (название цифровой электронной машины) и далее к латинскому *computare* (*считать*, *вычислять*) и указывает на специфику направления исследования. Компьютерная лингвистика разрабатывает возможные компьютерные подходы как к лингвистике в целом, так и отдельным лингвистическим вопросам. Одной из главных задач этой науки является моделирование искусственного языка.

Потребность в компьютерной лингвистике возникла во второй половине XX века в связи с развитием компьютерных технологий. Так, в 1950-х гг. в США были предприняты попытки использовать компьютеры для автоматического перевода текстов, в частности, интерес вызывал перевод русскоязычных научных изданий на английский язык.

Уже первые компьютеры производили разнообразные математические вычисления во много раз быстрее и правильнее, чем это мог делать человек. Отсюда и возникла идея возможности использовать компьютерные технологии для обработки языка. Кроме вопросов перевода, компьютерная лингвистика ставила задачей использование вычислительных и количественных методов для реконструкции ранних форм языков, т.е. продолжение работы в направлении сравнительно-исторического исследования.

В качестве даты рождения компьютерной лингвистики иногда называют январь 1954 года, когда в Джорджтаунском университете (США) был проведен «первый в мире публичный эксперимент по машинному переводу.

В то же самое времена под руководством крупнейшего математика и кибернетика Алексея Ляпунова начались активные работы по машинному переводу и в Москве. В созданную Ляпуновым группу вошли, в частности, тогдашние студенты и аспиранты, будущие «родители» отечественной компьютерной лингвистики Игорь Мельчук и Ольга Кулагина» [12, с. 35].

В 1962 году была основана ***Ассоциация машинного перевода и компьютерной л***ингвистики, которая в 1968 году получила название – ***«Ассоциация компьютерной лингвистики»*** (ACL). В состав ассоциации входят три отделения:

1) европейское (EACL);

2) североамериканское (NAACL);

3) азиатское (AACL).

Ассоциация ставит целью научное изучение языка с точки зрения вычислений. Компьютерные лингвисты разрабатывают вычислительные модели представления различных видов языковых явлений.

В числе первых методов, используемых компьютерной лингвистикой, были *лексикостатика* и *глоттохронология*. *Лексикостатикой* называется один из методов сравнительной лингвистики, применяемый для сравнения процентного соотношения родственных слов разных языков и определения взаимосвязи между словами и языками. Историческая лексикология описывает группы слов, которые были унаследованы из праязыка и имеют общее происхождение. Изменения в более поздних языках могли приводить к тому, что значительно менялось звучание и/или значение слова. Родственные связи в итоге становились малозаметными или вовсе незаметными. Компьютерные технологии должны были помочь в восстановлении таких связей и установлении *этимонов* (от греч. *etymon*), т.е. первоначального значения и формы слова.

Кроме того, компьютерные методы пытались использовать для выявления так называемых ʻложных родственниковʼ, т.е. слов, обладающих внешней схожестью в разных языках, но не имеющих никаких генетических связей. В качестве примера ʻложных родственниковʼ можно привести два схожих глагола: латинский *habēre* и немецкий *haben*. Обладая явной фонетической схожестью и общим лексическим значением (оба слова означают ʻиметьʼ, они, тем не менее, восходят к разным индоевропейским корням. Латинский глагол *habēre* восходит к лексеме ʻ$g^hab^hʼ$ (ʻдавать, получатьʼ). Родственными ему являются английское *give* и немецкое *geb*en. Немецкий глагол *habe* родственен английскому *have* и латинскому *capere* (ʻсхватывать, захватыватьʼ).

Глоттохронология (от греч. *glótta – язык, chrónos – время, lógos – слово, учение*) – это «область сравнительно-исторического языкознания, занимающаяся выявлением скорости языковых изменений и определением на этом основании времени разделения родственных языков и степени близости между ними» [28, с. 109]. Наиболее продуктивными для глоттохронологии

как раз и являются данные лексикостатистических исследований. Именно они позволяют определить «время разделения родственных языков, исходя из предположения об одинаковой скорости изменения той основной части словаря, которая нужна для обслуживания наиболее часто встречающихся и существенных ситуаций общения. К этой части словаря естественных языков принадлежат, согласно глоттохронологии принадлежат такие наиболее сохранные слова, как личные и вопросительные местоимения, некоторые глаголы, обозначающие движения (ʻприходитьʼ), элементарные физиологические функции и ощущения (ʻпитьʼ, ʻслышатьʼ, ʻвидетьʼ), обозначения размеров (ʻширокийʼ, ʻдлинныйʼ), космических явлений (ʻсолнцеʼ, ʻнебоʼ), животных (ʻчервьʼ, ʻзмеяʼ), цвета (ʻчерныйʼ) названия родства и т. п.» [28, с. 109]. Данная группа включает примерно от 100 до 200 слов. Если удается проследить их динамку во времени, то можно увидеть следующую закономерность: на протяжении тысячелетия (иногда нескольких тысячелетий) сохраняется в среднем не менее 80% словаря.

2. Межпредметные связи компьютерной лингвистики

В силу своей природы компьютерная лингвистика имеет весьма широкие межпредметные связи. Она естественным образом связана с различными разделами языкознания, частично пересекаясь «с обработкой естественных языков. Однако обработка естественных языков акцентирует внимание не на абстрактные модели, а на прикладные методы описания и обработки языка для компьютерных систем. Полем деятельности компьютерных лингвистов является разработка алгоритмов и прикладных программ для обработки языковой информации» [12, с. 45].

Из лингвистических направлений особо следует отметить структурализм, оказавший значительное влияние на развитие компьютерной лингвистики. Работы Р. Якобсона и его последователей были продуктивно использованы компьютерной лингвистикой. В середине XX века большую популярность приобретает математический структурализм, сблизивший математику и языкознание. Лингвистика, прежде сугубо гуманитарная наука, продемонстрировала способность к превращению в логически строгую дисциплину.

В число смежных дисциплин компьютерной лингвистики входят: *информатика, теория искусственного интеллекта*, *математика*, *логика*, *философия*, *когнитивистика* в целом и *когнитивная психология* в частности, *психолингвистика*, *антропология*, *нейробиология*.

Информатика (фр. *informatique*; англ. *computer science*) – наука о методах и процессах сбора, хранения, обработки, передачи, анализа и оценки информации с применением компьютерных технологий, обеспечивающих возможность её использования для принятия решений.

Особо следует отметить связь компьютерной лингвистики с теорией *искусственного интеллекта*. Авторство термина ʻискусственный интеллектʼ (англ. ʻ*artificial intelligence*ʼ) и его первое определение принадлежит

американскому ученому Джону Маккарти (1927–2011). Следует учесть, что английское слово ʻ*intelligence*ʼ означает ʻ*умение рассуждать разумно*ʼ, тогда как русские толковые словари обычно определяют понятие ʻ*интеллект*ʼ как «мыслительные способности человека, разум, уровень умственного развития». Русскому существительному в этом смысле соответствует английское ʻ *intellect*ʼ.

В 1956 году на семинаре в Дартмутском университете Дж. Маккарти охарактеризовал искусственный интеллект как «свойство искусственных интеллектуальных систем выполнять творческие функции, которые традиционно считаются прерогативой человека». Комментируя это понятие, исследователь уточняет: «Проблема состоит в том, что пока мы не можем в целом определить, какие вычислительные процедуры мы хотим называть интеллектуальными. Мы понимаем некоторые механизмы интеллекта и не понимаем остальные. Поэтому под интеллектом в пределах этой науки понимается только вычислительная составляющая способности достигать целей в мире» [«What is Artificial Intelligence?». Архивная копия сот 18 ноября 2015 на Wayback Machine FAQ от Джона Маккарти, 2007].

В современной научной литературе можно найти разные дефиниции понятия ʻискусственный интеллектʼ:

– наука и технология создания интеллектуальных машин, особенно интеллектуальных компьютерных программ;

– свойство интеллектуальных систем выполнять творческие функции, которые традиционно считаются прерогативой человека.

Как научное направление, теория искусственного интеллекта решает задачи аппаратного или программного моделирования таких видов человеческой деятельности, которые традиционно считаются интеллектуальными. Естественно, что человеческий язык, будучи средством познания мира, средством мышления, играет здесь ведущую роль. Создание искусственного интеллекта необходимо опирается на языковые факторы. Разумные рассуждения и действия, воссоздаваемые с помощью специальных программ и вычислительных устройств, являются объектом исследования информатики. Здесь структура интеллектуальной система включает три основных блока:

1) базу знаний;

2) так называемый решатель;

3) интерфейс (от англ. *interface*), который должен обеспечить процесс общения с электронным вычислительным устройством.

Предметом *психолингвистики* является взаимосвязь между языковыми факторами и психологическими аспектами. Она выявляет и изучает, «процессы речеобразования, а также восприятия и формирования речи в их соотнесенности с системой языка» [28, с. 404].

Логика, основателем которой является античный ученый Аристотель, сегодня представляет собой нормативную науку о законах, формах и приемах интеллектуальной деятельности

3. Теоретическая и прикладная компьютерная лингвистика

Современная компьютерная лингвистика представлена *теоретическим* и *прикладным* направлениями. Объектом внимания теоретического направления являются вопросы общей теоретической лингвистики и когнитивистики. Это направление осуществляет разработки в области формальных теорий грамматики, в частности синтаксического анализа, и семантики.

Одним из классиков теоретической компьютерной лингвистики является английский ученый *Алан Мэтисон Тьюринг* (1912–1954), который работал в области математики, логики и криптографии. *Криптография* (от греч. *kryptos – тайный*, *скрытый* + *пишу*, букв. *тайнопись*) занимается проблемами обеспечения конфиденциальности, проверки подлинности информации, вопросами шифрования и т. п. В 1936 году А. Тьюрингом была предложена абстрактная вычислительная машина, получившая название «Машина Тьюринга». Это своего рода модель компьютера общего назначения, т.е. компьютера, который способен решать разнообразные задачи, выраженные в виде программы. Благодаря разработке А. Тьюринга было формализовано понятие алгоритма, а его идеи до сих пор плодотворно используются при проведении теоретических исследований и решении практических задач компьютерной лингвистики. Труды А. Тьюринга внесли значительный вклад в развитие информатики и теории искусственного интеллекта.

Теоретическая компьютерная лингвистика опирается формальную логику и так называемые *символические* подходы. Предметом математической формальной логики являются:

– теория моделей;

– теория доказательств;

– теория множеств;

– теория рекурсии (теория вычислимости).

Символический подход связан с понятием *символического искусственного интеллекта*. Этот термин используется для обозначения совокупности методов в исследованиях искусственного интеллекта, основанных на высокоуровневых представлениях проблем.

Прикладная компьютерная лингвистика занимается непосредственно практическими проблемами моделирования языка. В сфере ее внимания находятся нейронные сети, т.е. цепи биологических нейронов. Современные исследования преимущественно направлены на искусственные нейронные сети, которые, в свою очередь, используются для решения задач искусственного интеллекта. Биологические связи моделируются в искусственных нейронных сетях. Эти связи представляются в виде веса между узлами, когда положительный вес отражает возбуждающую связь, а отрицательный – тормозную.

Задачами прикладной компьютерной лингвистики являются:

1) обработка естественного языка (англ, *natural language processing*; синтаксический, морфологический, семантический анализы текста);

2) создание электронных словарей, тезаурусов, онтологий (например, Lingvo). Словари используют для автоматического и автоматизированного переводов, проверки орфографии и т.д.;

3) автоматический перевод текстов посредством специализированных программ;

4) автоматическое извлечение фактов из текста (извлечение информации; англ. *fact extraction, text mining)*;

5) автореферирование (англ, *automatic text summarization).* Эта функция включена, например, в *Microsoft Word*;

6) построение систем управления знаниями (экспертные системы);

7) создание вопросно-ответных систем (англ, question answering systems);

8) оптическое распознавание символов (англ. *OCR). н*апример, программа FineReader;

9) автоматическое распознавание и синтез речи.

Помимо этого, в сферу интересов компьютерной лингвистики входят:

– компьютерный анализ жанра и характеристик текста (который является еще более сложным, чем просто анализ сюжета);

– компьютерный анализ блогосферы (как вариант анализа корпуса текстов);

– создание так называемой семантической паутины Интернета (формирование пространства знаний) и поиск этих знаний в ней.

МОДУЛЬ II

Тема 1
КОРПУСНАЯ ЛИНГВИСТИКА

План:

1. Языковой корпус и корпусная лингвистика.
2. Современные корпусы национальных языков.
3. Национальный русского языка.
4. Белорусский N-корпус.

Ключевые понятия: *корпусная лингвистика*, *языковой корпус, национальный корпус*.

1. Языковой корпус и корпусная лингвистика

В 1960-ые годы в Университете Брауна (*Brown University*) был создан первый большой компьютерный *лингвистический корпус*, который получил название, соответственно, Брауновского корпуса (*Brown Corpus*, BC). Лингвистическим, или языковым, корпусом называют «большой,

представленный в машиночитаемом формате, унифицированный, структурированный, размеченный филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [15, с. 10]. Лингвистическим корпусом также называют совокупность текстов, собранных в соответствии с определенными принципами и размеченных по определенному стандарту. Такие тексты внутри корпуса должны быть обеспечены специализированной поисковой системой, в них можно найти типичное употребление отдельного слова или фразы, а также значение и грамматическую функцию.

В понятие «корпус текстов» входит также система управления текстовыми и лингвистическими данными, которую называют *корпусным менеджером* (или корпус-менеджером) (англ. *corpus manager*). Это специализированная поисковая система, включающая в себя программные средства для поиска запрашиваемых данных в корпусе и предоставления их пользователю в удобной форме, а также для получения статистической информации [15, с. 12].

Еще одной особенностью корпуса является возможность построения конкорданса – списка всех употреблений данного слова в контексте со ссылками на источник.

Современные национальные языковые корпуса – это информационно-справочные системы, основу которых составляют собрания текстов в электронной форме. Национальные корпуса представляют конкретный язык на определенном этапе (этапах) существования и отражают многообразие жанров, стилей, территориальных и социальных вариантов.

Представление систематизированных текстов в корпусном варианте позволяет решить ряд задач. Во-первых, языковые средства (слова, конструкции) подаются в реальном контексте их употребления; во-вторых, материал является представительным (особенно при больших объемах корпуса); в-третьих, корпусные данные используются многократно для проведения самых разных исследований.

Брауновский корпус включал в себя 500 фрагментов текстов по 2 тысячи слов каждый. Благодаря ему был задан стандарт в 1 миллион словоупотреблений, на который ориентировались в дальнейшем при создании корпусов на других языках. Данный объем позволяет представить, в основном, наиболее частотные слова. Лексемы и грамматические конструкции средней частоты употребления встречаются только по нескольку раз на миллион словоупотреблений. В качестве примера можно привести английские лексемы *polite* (*вежливый*) и *sunshine* (*солнечный свет*), встретившиеся по семь раз, или конструкцию *polite letter* (вежливое письмо), употребленное только один раз. При этом выражение *polite conversation* (*вежливый разговор*), слова *smile* (*улыбка*), *request* (запрос) вообще оказались вне корпуса.

В 1970-ые годы коллективом сотрудников филологического факультета Ленинградского государственного университета имени А.А. Жданова

и сотрудников Лаборатории семиотики Научно-исследовательского института прикладной математики и кибернетики при Горьковском государственном университете им. Н. А. Лобачевского по схожей с Брауновским корпусом модели был разработан частотный словарь русского языка, включавший около 40 000 слов. В 1977 году это лексикографическое издание объемом 935 страниц вышло из печати под редакцией Л. Н. Засориной. В основу словаря легли тексты, относящиеся к четырем функционально-речевым сферам:

1) художественной прозе (25,4%);
2) драматургии (27,2%);
3) науке (23,6%);
4) публицистике (23,8%).

Каждая группа представлена приблизительно 250 000 словоупотреблений. Отдельное внимание к драматургии авторы объясняют сложностью отражения разговорной речи: «Не располагая записями разговоров в достаточном количестве, мы использовали в качестве источников разговорной речи современную реалистическую драматургию. Элементы разговорного словоупотребления, безусловно, проникают также в публицистические тексты, особенно массового назначения, а также литературно-художественные произведения, чаще – повести и рассказы» [59].

Словарь состоит из трех частей: Алфавитно-частотного словника, Частотного словника, Статистической структуры словаря. Первая часть является наиболее важной и объемной: она включает в себя все лексемы, встретившиеся в текстах (39 268 слов: от ‘а’ до ‘ящичный’). Лексемы сопровождаются несколькими количественными характеристиками: указывается их частота с детализацией (общая частота по всей выборке, частота по подвыборкам, т.е. по каждой из четырех жанровых групп текстов) и количество текстов по жанрам, в которых встретилось данное слово. Алфавитно-частотный словник дает сведения о функциональной отнесенности слова.

В Частотный словник включены слова с частотой 10 и выше. Авторы в предисловии указывают: наибольшую частоту имеет предлог *в*(*во*) – 42 854. Они отмечают, что частые слова составляют 23,02% всего словника, но покрывают 92,4% всего текста, тогда как остальные 30 224 слова покрывают только 7,6% всей выборки.

В 1980-е гг. в Швеции в университете города Уппсала был создан русский корпус, построенный по аналогичной частотному словарю модели.

В 1985 году по инициативе академика А. П. Ершова был создан отдел машинного фонда русского языка. В создании самого Машинного фонда в период с 1986 по 1990 годы принимали участие более 40 организаций.

С возникновением языковых корпусов стал формироваться и специальный раздел языкознания, который занялся проблемами разработки, создания и использования текстовых корпусов. Данное направление получило название корпусной лингвистики и в XXI столетии активно разрабатывается.

2. Современные корпусы национальных языков

Современное состояние компьютерной техники и развитие компьютерной лингвистики позволили создать или, по крайнее мере, начать активную разработку целого ряда корпусов разных языков.

Английский язык представлен *в Британском национальном к*орпусе (*BNC* от англ. *British National Corpus*), в котором отражены 100 миллионов слов и образцы их употребления в письменном и разговорном британском английском языке конца XX века. Всемирный английский в корпусе не представлен. Корпус создавался в период с 1991 по 1994 гг., после чего новые примеры употребления в него не добавлялись, но были внесены незначительные изменения.

Примерно 90% Британского национального корпуса составляют образцы употребления письменного языка. Источниками послужили общенациональные и региональные СМИ, научные журналы, разнообразная периодика, художественная литература, а также ряд неопубликованных материалов (письма, студенческие эссе, сценарии, речи и др.). На разговорный раздел корпуса приходится 10% от общего объема, он состоит из двух частей, каждую из которых составляют образцы разговорного языка, записанные с помощью практической транскрипции.

Первая часть, которую называют демографической, основана на записях спонтанных разговоров, происходивших в реальных условиях. Для организации таких записей привлекались волонтеры из различных возрастных групп, регионов и социальных слоев. Разговоры могли быть записаны в разнообразных ситуациях: деловые встречи, обсуждения в радиопередачах, телефонное общение и т.д.

Вторая часть основана на контекстно-зависимых образцах. Это записи, сделанные в ходе особых встреч, мероприятий.

В 1960-е гг. в городе Мангейм началась работа над созданием национального корпуса немецкого языка. Этот корпус часто называют Мангеймским. Другие названия: *German Reference Corpus, IDS corpora, COSMAS corpora.* В 2004 году корпус получил официальное название – *Deutsches Referenzkorpus (DeReKo*). В основной корпус включены полные лицензированные тексты (не фрагменты), которые создавались с 1956 года, причем их база постоянно пополняется. Имеется также исторический корпус, корпус новых поступлений и подкорпус разговорной речи.

В Китае реализуется ряд проектов в области корпусной лингвистики китайского языка. Их задачей, соответственно, является создание специализированных корпусов. В числе общедоступных можно назвать *Корпус современного китайского языка* (*The Modern Chinese Language Corpus*), создаваемый в Центре китайской лингвистике при Пекинском университете. Корпус имеет ряд особенностей, частично обусловленных спецификой представляемого языка. «Метод поиска в данном корпусе базируется на фактическом расстоянии между иероглифами/словами и позволяет построить

конкорданс – список всех употреблений данной словоформы в контексте. В Корпусе современного китайского языка имеется только метаразметка, при этом отсутствуют такие обязательные признаки корпуса, как морфологическая и синтаксическая разметка, поэтому в строгом смысле эту базу нужно рассматривать не как корпус языка, а скорее как текстовый архив. Неаннотированность делает использование данного корпуса в области грамматических исследований затруднительным» [21].

*Сбалансированный корпус китайского яз*ыка оценивается как значительный шаг в развитии китайской корпусной лингвистики. Он включает в себя два подкорпуса: первый состоит из почти 13 миллионов словоупотреблений современного китайского языка, второй содержит около 100 миллионов иероглифов древних текстов.

Национальный корпус польского языка имеет название Narodowy Korpus Języka Polskiego (NKJP) и достаточно широко представляет этот язык в разных его проявлениях. Из корпусов славянских языков также следует выделить Чешский национальный корпус, созданный в Карловом университете Праги.

3. Национальный корпус русского языка

Национальный корпус русского (НКРЯ), как говорится, на сайте https://ruscorpora.ru/, – это представительная коллекция текстов на русском языке общим объемом более 1,5 миллиарда слов, оснащенная лингвистической разметкой и инструментами поиска. Корпус открыт 29 апреля 2004 года. Сегодня он представляет собой «собрание независимых корпусов, каждый из которых предназначен для решения определенных лингвистических задач. Каждая из этих коллекций текстов является большой по объёму и представительной, что делает их ценным материалом для количественных и качественных исследований» [61].

В корпус включены письменные тексты и записи устных речей. Письменные тексты отражают такие сферы, как:

– художественная литература,
– мемуаристика,
– наука,
– религия,
– повседневная печатная продукция.

Каждый из перечисленных подкорпусов имеет свои особенности. Так, корпус «Русская классика» позволяет максимально полно представить наследие русской классики без каких-либо ограничений, например, найти черновики или редакционные варианты произведений.

Специфика лингвистических задач определяет состав корпуса и разметку, которая в нем используется. Например, поэтический корпус может служить основой для стиховедческих исследований, поэтому в нем есть особая разметка, связанная с ключевыми для стиховедения понятиями –

метром и ритмом. Вообще наличие системы разметок отличает корпус от других так называемых простых коллекций (библиотек) текстов. «Разметка корпуса (*tagging, annotation*) заключается в приписывании текстам и их компонентам специальных тегов: собственно *лингвистических*, описывающих лексические, грамматические и прочие характеристики элементов текста, и внешних, *экстралингвистических* (сведений об авторе и о тексте: автор, название, год и место издания, жанр, тематика)» [15, с. 34]. *Тег*, либо *тэг* (от англ. tag: метка, бирка, ярлык) – ассоциированное ключевое слово, относящееся к какой-либо информации (текст, фото, видео, закладки браузера и другие файлы). Такие метаданные помогают полнее описать эти куски информации и быстро находить их через поисковый запрос. Тэги используются без жёстких правил автором или конечным потребителем.

Научная, познавательная, учебная ценность корпуса определяется разнообразием разметки. Национальный корпус русского языка на данный момент предлагает четыре типа разметки: метатекстовую, морфологическую, акцентную и семантическую. Нужно подчеркнуть, что система разметок постоянно совершенствуется.

Кроме того, есть подкорпусы поэтических текстов, диалектных текстов и корпусы параллельных текстов. Поэтический корпус является одним из самых сложно устроенных. В нем представлены только те поэтические тексты, которые размечены вручную. Поиск в поэтическом корпусе может осуществляться по разным критериям: по стихотворному размеру (хорей, дактиль, ямб и др.), по числу стоп, по клаузуле (женская или мужская), строфике (одическая, онегинская строфа, рондо и т.д.).

В Национальном корпусе есть сегменты, предназначенные для специалистов узкого профиля. Так, например, специалистам в области истории языка будет интересен корпус берестяных грамот, включающий оригинальные тексты грамот и их переводы на современные русский и английский языки, корпусы древнерусского, старорусского и церковнославянского языков.

Синтаксический корпус предоставляет возможность рассмотрения синтаксических деревьев – систем связей между синтаксическими единицами.

Национальный корпус русского языка дополняется новыми разделами, в числе которых подкорпус социальных сетей, призванный помочь в изучении языка, используемого в интернет-общении.

Особый раздел представляет параллельный корпус, содержащий тексты на двух языках. Корпусы, параллельные с русским, доступны для следующих языков: английский, армянский, белорусский, болгарский, бурятский, испанский, итальянский, китайский, латышский, немецкий, польский, украинский, французский, шведский, эстонский и многоязычный. Есть отдельный газетный корпус, содержащий материалы СМИ начала XXI века. Двуязычные корпусы являются не собранием переводов, а системой

возможностей работы с языковым материалом. Исследователь может поставить разные задачи, например:

– определить соответствия между системами русских и английских модальных глаголов;

– определить русские соответствия английскому слову *cat* – кот или кошка и т.д.

Для упрощения системы поиска на главной странице размещена строка «Обзор возможностей», которая позволяет путем введения слова или фразу найти разнообразные виды поисковой выдачи, имеющиеся в корпусе.

4. Белорусский N-корпус

Белорусский язык представлен в белорусском N-корпусе, который сейчас находится в стадии активной доработки. После соответствующих юридических процедур он получит статус и название национального корпуса.

По состоянию на начало 20-ых гг. XXI века корпус включал ряд разделов, в числе которых:

– корпус текстов современного белорусского языка со структурной и грамматической разметкой, а также с их паспортизацией;

– русско-белорусский параллельный корпус, в состав которого пока входят только переводы кодексов Республики Беларусь;

– Библейский корпус, вмещающий 16 переводов Библии на белорусский язык, а также тексты на других языках (латинском, еврейском, польском и др.) и позволяющий сравнивать тексты переводов, находить нужные языковые элементы и т.д.

Корпус современного белорусского языка включает несколько подкорпусов:

– основной,

– неразобранных текстов,

– газет

– сайтов.

Общий объем этого корпуса уже составляет около 177 миллионов словоупотреблений, а в совокупности с неразобранными текстами – около 1,07 миллиарда словоупотреблений.

В Интернете постоянно появляется информация о новых разделах и возможностях белорусского N-корпуса. Создатели заявили о существенном обновлении интерфейса, расширении грамматической базы, дополнении информации об источниках слова и многом др.

Сегодня белорусский N-корпус предоставляет возможности проверки правописания в разных браузерах, он содержит фонетический онлайновый конвертер, позволяющий трансформировать тексты в транскрипцию с использованием Международного фонетического алфавита (IPA) и белорусской школьной транскпирции и т.д.

В тестовом варианте находятся проекты, созданные на основе грамматической базы, которая представляет собой собрание лексем с морфологическими и другими пометами. База корпуса является актуальной, в том числе и потому, что содержит слова разного типа: как уже зафиксированные в нормативных словарях, так и появившиеся недавно и не получившие словарной прописки.

Белорусская корпусная лингвистика относится к числу наиболее востребованных и практически значимых направлений исследований.

Тема 2
СОВРЕМЕННЫЕ ПОИСКОВЫЕ СИСТЕМЫ

План:

1. Поиск информации в Интернете. Поисковые системы.
2. Информационно-поисковые тезаурусы.
3. Лингвистические особенности дескрипторов.

Ключевые понятия: *поисковая система, тезаурус, словарь, дескриптор, аскриптор*

1. Поиск информации в Интернете. Поисковые системы

В условиях избытка информации, находящейся в Интернете, важнейшей задачей является поиск того, что необходимо конкретному пользователю. Поиск – это определенная последовательность действий, операций, целью которых является сбор и обработка необходимой информации. Поиск информации в Интернете осуществляется за счет использования комплекса программ, именуемых поисковой машиной. Такая машина работает в составе поисковой системы – «автоматизированного программно-аппаратного комплекса с веб-интерфейсом, предоставляющего возможность поиска информации в Интернете» [12, с. 27.].

Название ‘веб-интерфейс’ происходит от английских двух слов, первое из которых w*eb* (букв. ‘паутина’) является «частью сложных слов, обозначающей: относящийся к сети Интернет» [40, с. 76]; второе *interface* (от лат. *inter – между, посередине* и *face* – лицо) – «система унифицированных связей и сигналов, посредством которых устройства вычислительной системы взаимодействуют друг с другом» [40, с. 158].

Сегодня существует достаточно много поисковых систем, самой известной из которых является *Google* [63], самой популярной в русскоязычном сегменте Интернета – *Яндекс* [67], а самой старой – *Yahoo* [сайт]. Поиск необходимой информации может осуществляться разными способами. Для того, чтобы сделать его эффективным, пользователь знать «либо адрес её местоположения (например, адрес *Ыт*\-страницы или файла), либо пользователя Интернета, который может предоставить информацию. Если мы

не знаем ни адреса, ни человека, который мог бы нам помочь, то следует перейти к вопросам *«Как можно узнать адрес размещения информации?»* или *«Как найти человека, который мог бы нам помочь с поиском информации?»* [12, с. 27.].

В справочной литературе перечисляются четыре этапа поиска информации:

1) определение (уточнение) информационной потребности и формулировка информационного запроса;

2) определение совокупности возможных держателей информационных массивов (источников);

3) извлечение информации из выявленных информационных массивов;

4) ознакомление с полученной информацией и оценка результатов поиска [12, с. 27.].

Как уже отмечалось, современное интернет-пространство предлагает большое количество систем поиска необходимой пользователю информации. Крупнейшей из них является *Google* (www.google.com): на ее долю приходится более 77% запросов. «Из-за популярности поисковой системы в английском языке появился неологизм to google или to Google (аналог в русском компьютерном сленге – гуглить), использующийся для обозначения поиска информации в Интернете с помощью Google. Именно с таким определением глагол занесён в наиболее авторитетные словари английского языка – Оксфордский словарь английского языка и Merriam-Webster, хотя в других источниках приводятся примеры его использования для обозначения поиска вообще чего-либо в Интернете» [12, 2016, с. 63]. Поиск в системе Google может применяться для документов разных форматов: PDF, RTF, PostScript, Microsoft Word, Microsoft Excel, Microsoft PowerPoint и других.

Второй по популярности является поисковая система *Yahoo* (www.yahoo.com), которая предоставляет своим пользователям ряд сервисов, включая электронную почту Yahoo! Mail.

В русскоязычной части Интернета большой популярностью пользуется система *Яндекс* (www.yandex.ru). Современный Яндекс – это не только поисковик, это мультипортал, предлагающий более 50 сервисов, среди которых Яндекс. Поиск, Яндекс. Карты, Яндекс. Маркет и ряд других. Авторство бренда принадлежит Аркадию Воложу и Илье Сегаловичу, которые заменили первую букву в лексеме ‘index’.

Премиальный медийносервисный портал *Рамблер* (www.rambler.ru) существует с 1996 года и с 2012 специализируется на специализированных новостях. Его название восходит к английскому слову ‘Rambler’ – *странник, бродяга*. Сегодня *Рамблер* – это «индивидуальная картина дня и помощь в главных аспектах жизни. На «Рамблере» можно прочитать важные новости, разобраться в сфере финансов, недвижимости и авто, отправиться в путешествие, посмотреть популярные видео, купить билеты в кино или театр, собрать ребенка в садик и школу, познакомиться, узнать точный

прогноз погоды и весело провести выходные. «Рамблер» – портал, которому доверяют» [12, с. 65].

Российская система *Nigma* (Нигма. РФ) (www.nigma.ru) получила именование по названию рода пауков *Nigma walckenaeri*, что объясняется стремлением номинаторов подчеркнуть связь с Интернетом как Всемирной паутиной.

Из числа других систем следует отметить *Bing* (www.bing.com), занимающую второе место по объему трафика и обладающую рядом эксклюзивных возможностей, в числе которых «просмотр результатов поиска на одной странице (вместо пролистывания многочисленных страниц результатов поиска), а также динамическое корректирование объёма информации, отображаемой для каждого результата поиска (например, только название, краткая или большая сводка)» [12, с. 65].

В китайском сегменте Интернета лидером по числу запросов является поисковая система *Baidu* (www.baidu.com). Энциклопедия Байду (Байдупедия) превосходит в этом отношении китайскую Википедию.

2. Информационно-поисковые тезаурусы

Термин ‘тезаурус’ сегодня используется в нескольких значениях. Для информационной лингвистики актуальны следующие дефиниции тезаурусов:

1) информационно-поисковый ресурс, описывающий отношения между терминами предметной области (такие тезаурусы создаются экспертами в определенной предметной области и предназначаются для помощи при информационном поиске);

2) словарь, в котором слова и словосочетания с близкими значениями сгруппированы в единицы, называемые понятиями, концептами или дескрипторами, и в котором явно (в виде отношений, иерархии) указываются семантические отношения между этими понятиями (концептами, дескрипторами);

3) недавно появившиеся лингвистические ресурсы типа WordNet и EuroWordNet, описывающие отношения между лексическими значениями естественного языка как иерархическую систему групп синонимов – синсетов.

Синсетом (от английского “synset” – сокращение от “synonym set”) называют один из ключевых элементов в лексической базе данных WordNet и подобных ей системах. Это понятие играет важную роль в компьютерной лингвистике и обработке естественного языка. Каждый синсет содержит несколько лексических единиц (слов или словосочетаний). К синсету прилагается краткое определение, называемое глоссой, которое объясняет значение данного набора синонимов. Синсеты могут также включать примеры использования слов в контексте.

В информатике используется также термин ‘ассоциативный тезаурус’, который употребляется для ссылки на ресурсы, автоматически создаваемые на основе обработки корпусов и показывающие совместную встречаемость пар слов в документах.

Информационно-поисковые тезаурусы появились в 60-е годы 20 века. В это время большинство информационно-поисковых систем не являлись полнотекстовыми, а хранили достаточно ограниченный набор информации о документе: библиографические данные, реферат. Добавление списка ключевых слов, характеризующих основное содержание документа, существенно расширяли возможности поиска документов. С начала семидесятых годов создаются национальные и международные стандарты разработки информационно-поисковых тезаурусов. В соответствии с определениями стандартов информационно-поисковый тезаурус – это нормативный словарь, явно указывающий отношения между терминами и предназначенный для описания содержания документов и поисковых запросов.

Основными целями разработки информационно-поисковых тезаурусов являются:

– обеспечение перевода естественного языка документов и пользователей на один и тот же словарь, используемый для индексирования и поиска, таким образом, различия в лексическом составе документа и запроса пользователя сводились к одним и тем же единицам тезауруса;

– обеспечение последовательного использования единиц индексирования;

– обеспечение отношений между терминами - отношения между единицами тезауруса позволяют найти оптимальный термин для описания документа или запроса;

– использование как поискового средства при поиске документов [31, с. 23].

Важным элементом тезауруса являются *дескрипторы* (от лат. *descriptor* – «описывающий»). Дескриптор – это лексическая единица (слово, словосочетание) информационно-поискового языка, служащая для описания основного смыслового содержания документа или формулировки запроса при поиске документа (информации) в информационно-поисковой системе. Дескриптор однозначно ставится в соответствие группе ключевых слов естественного языка, отобранных из текста, относящегося к определённой области знаний.

Каждый дескриптор, внесенный в тезаурус, представляет собой отдельное понятие той или иной области. Он может быть однословным или многословным, причем во многих тезаурусах особое внимание уделяется как раз многословным описаниям.

Набор дескрипторов тезауруса должен удовлетворять следующим требованиям:

– посредством выделенных дескрипторов должно быть возможно описать темы абсолютного большинства текстов предметной области;

– для уменьшения субъективности индексирования множество дескрипторов не должно включать совокупности близких дескрипторов, формируются классы условной эквивалентности, когда совокупности близких, но различных понятий сводятся к одному дескриптору;

– дескриптор должен быть сформулирован однозначно, его подразумеваемое в рамках тезауруса значение должно быть понятно пользователю. Если однозначный и ясный дескриптор подобрать не удается, термин, взятый в качестве дескриптора, снабжается *релятором* (краткой пометой) или комментарием.

Релятором называется символ или слово, которые используются для различения значений многозначного термина. Он не является независимой лексической единицей информационного поискового языка.

3. Лингвистические особенности дескрипторов

Создание дескрипторов требует соблюдения определенных лингвистических подходов, часть из которых зафиксирована в соответствующих ГОСТах.

Если опорным словом дескриптора является существительное, то словосочетание должно удовлетворять следующим условиям, которые предполагают, что:

– значение словосочетания не выводится из значений его компонентов, например, *ЧЕРНЫЙ ЯЩИК, АБСОЛЮТНО ЧЕРНОЕ ТЕЛО, ЦАРСКАЯ ВОДКА*;

– хотя бы один из компонентов словосочетания не употребляется в составе других сочетаний или употребляется всегда в другом смысле, например, *ТОРГОВЛЯ НА ВЫНОС, ЛЕГКАЯ ПРОМЫШЛЕННОСТЬ*;

– для данного словосочетания в словнике тезауруса существуют полные синонимы, например, *НАТРИЯ ХЛОРИД = ПОВАРЕННАЯ СОЛЬ*;

– данное словосочетание является устойчивым словосочетанием с именем собственным: *ТАБЛИЦА МЕНДЕЛЕЕВА*, *ЗАКОН БОЙЛЯ-МАРИОТТА*;

– отдельные слова словосочетания имеют слишком широкое значение, например, слово ʻмашиныʼ в словосочетаниях: *СТРОИТЕЛЬНЫЕ МАШИНЫ, ЭЛЕКТРИЧЕСКИЕ МАШИНЫ*;

– для данного словосочетания в словнике тезауруса существует общепринятая аббревиатура, например: *ПОВЕРХНОСТНО-АКТИВНЫЕ ВЕЩЕСТВА – ПАВ, УНИВЕРСАЛЬНАЯ ДЕСЯТИЧНАЯ КЛАССИФИКАЦИЯ – УДК, ИНФОРМАЦИОННО-ПОИСКОВЫЙ ТЕЗАУРУС – ИПТ, ЭЛЕКТРОННО-ВЫЧИСЛИТЕЛЬНАЯ МАШИНА – ЭВМ*;

– разбиение словосочетаний на отдельные компоненты приводит к потере важных для поиска семантических связей. Так, разбиение языкового выражения ʻязык программированияʼ не позволяет установить связи с такими языковыми выражениями, как «АЛГОЛ», «КОБОЛ», «ФОРТРАН».

Многие понятия могут выражаться с помощью двух разных или большего количества терминов, один из которых является основным. В таких случаях разные термины должны быть эквивалентны между собой, что достигается за счет использования:

– собственно синонимов;

– лексических вариантов;

– квазисинонимов.

Лексические варианты отличаются от синонимов тем, что они представляют собой некоторую модификацию одного и того же выражения, например, различное написание, аббревиатуры, и т.п.

Лексическая единица в информационно-поисковом тезаурусе, которая не может быть использована для координатного индексирования и подлежит замене одним или несколькими дескрипторами, имеет название *аскриптора*.

В качестве аскрипторов часто могут использоваться квазисинонимы, то есть такие термины, значения которых различаются, но рассматриваются как эквиваленты для целей тезауруса. Например, в качестве квазисинонимов могут выступать антонимы: *ядерная опасность – ядерная безопасность*. Еще одним видом квазисинонимов является случай, когда в качестве дескриптора рассматривается некий обобщающий тип, а его подвиды описываются как аскрипторы к этому дескриптору.

Аскрипторы, не совпадающие по значению, вводятся по ГОСТу в нескольких случаях. Например, когда они являются относительными синонимами (если случаи несовпадения значений несущественны для задач ИПТ):

СТОЛ = ДИЕТА = ПИТАНИЕ,

БЮРО = КОНТОРА = ФИРМА,

ВИНТ = БОЛТ.

Допускается установление эквивалентности также между единицами, различными по значению, но семантически связанными, в тех случаях, когда отождествление этих понятий полезно для функционирования информационной системы:

УСТОЙЧИВОСТЬ = НЕУСТОЙЧИВОСТЬ,

ТОРГОВЛЯ == ПРОДАЖА,

РЕКА = РУЧЕЙ,

МАСЛО = СМАЗКА.

Основными типами отношений, обычно отражаемых в информационно-поисковых тезаурусах являются следующие:

род – вид,

часть – целое,

причина – следствие,

сырье – продукт,

административная иерархия,

процесс – объект,

функциональное сходство,

процесс – субъект,

свойство – носитель свойства,

антонимия.

Список использованных источников

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебное пособие для студентов высших учебных заведений, обучающихся по направлению 231300 – «Прикладная математика» / [Большакова Е.И. и др.]; М-во образования и науки Российской Федерации, Московский гос. ин-т электроники и математики Москва: Московский гос. ин-т электроники и математики, 2011. – 272 с.
2. Автоматический анализ текстов: сб. науч. ст. / МГПИИЯ; редкол.: А.В. Зубов (отв. ред.) [и др.]. – Минск, 1976. – 198 с.
3. Актуальные проблемы компьютерной лингвистики: сб. науч. ст. / МГЛУ; редкол.: А.В. Зубов (отв. ред.) [и др.]. – Минск, 2005. – 318 с.
4. Анисимов, Д.В. Правда о машинном переводе / Д.В. Анисимов. – М.: Сам Полиграфист, 2014. – 340 с.
5. Баранов, А.Н. Введение в прикладную лингвистику: Учебное пособие / А.Н. Баранов. – М.: Эдиториал УРСС, 2001. – 360 с.
6. Бейтсон, Г. Экология разума: Избранные статьи по антропологии, психиатрии и эпистемологии / Г. Бейтсон / Пер. Д.Я. Федотова, М.П. Папуша; вступ. ст. А.М. Эткинда. – 1-е изд. – М.: Смысл, 2000. – 476 с.
7. Болховитянов, А.В. Алгоритмы морфологического анализа компьютерной лингвистики: учеб, пособие для студентов вузов, обучающихся по направлению 035000.62 – Издательское дело / А.В. Болховитянов, А.М. Чеповский; М-во образования и науки Российской Федерации, Федеральное гос. бюджетное образовательное учреждение высш. проф. образования Московский гос. ун-т печати им. Ивана Федорова. – Москва: МГУП, 2013. – 198 с.
8. Валипур, Али-Реза. Анализ и синтез глагольных форм и конструкций при машинном переводе с русского языка на персидский: дисс. канд. филол. наук / Али-Реза Валипур. – М.: МГУ, 1998. – 121 с.
9. Веденов, А.А. Моделирование элементов мышления / А.А. Веденов. – М.: Наука, 1988. – 159 с.
10. Вопросы общей и прикладной лингвистики: сб. науч. ст. / МГПИИЯ; редкол.: А.В. Зубов (отв. ред.) [и др.]. – Минск, 1975. – 297 с.
11. Горошко, Е.И. Лингвистика Интернета: формирование дисциплинарной парадигмы / Е.И. Горошко // Жанры и типы текста в научном и медийном дискурсе. – Орел: Картуш, 2007. – Вып.5. – С. 223–237.
12. Гуслякова, А. В. Информационные технологии и лингвистика XXI века / А.В. Гуслякова. – М.: «МГПУ, 2016. – 69 с.
13. Денисов, П.Н. Принципы моделирования языка / П.Н. Денисов. – М.: МГУ, 1965. – 151 с.
14. Ефимов, Н.Н. Основы информатики. Введение в искусственный интеллект / Н.Н. Ефимов, В.С. Фролов. – М.: МГУ, 1991. – 115 с.

15. Захаров, В.П. Корпусная лингвистика: Учебник для студентов направления лингвистика / В.П. Захаров, С.Ю. Богданова. – 3-е изд., перераб. и доп. – СПб.: Изд-во С-Петерб. ун-та, 2020. – 234 с.
16. Зелко, В.М. Проблемы разработки лингвистического обеспечения системы китайско-русского информационного машинного перевода: дисс. канд. филол. наук / В.М. Зелко. – М.: Ин-т языкознания АН СССР, 1991. –165 с.
17. Зубов, А. Б. Информационные технологии в лингвистике/ А.Б. Зубов, И.И. Зубова. М.: «Академия», – 2004. – 208 с.
18. Зубова, И.И. Информационные технологии в лингвистике / И.И. Зубова. – Минск: МГЛУ, 2001. – 211 с.
19. Караулов, Ю.Н. Методология лингвистического исследования и машинный фонд русского языка / Ю.Н. Караулов // Машинный фонд русского языка: идеи и суждения. – М.: Наука, 1986. – С. 13–25.
20. Коваль, С.А. Лингвистические проблемы компьютерной морфологии / С.А Коваль. – Санкт–Петербург: СПбГУ, 2005. – 147 с.
21. Колпачкова, Е.Н. Корпусы китайского языка: современное состояние и основные проблемы / Е.Н. Колпачкова // Корпусная лингвистика – 2015: Труды международной конференции. Отв. ред. В.П. Захаров, О.А. Митрофанова, М.В. Хохлова. – СПб.: Изд-во С-Петерб. ун-та, 2015. – С. 278 – 286.
22. Комиссаров, В.Н. Теория перевода (лингвистические аспекты): Учеб. для ин-тов и фак. иностр. яз. / В.Н. Комиссаров. – Репр. изд. – М.: Яльянс, 2103. – 250 с.
23. Компьютерная лингвистика и обучение языку: сб. науч. ст. / МГЛУ; редкол.: А.В. Зубов (отв. ред.) [и др.]. – Минск, 2000. – 219 с.
24. Компьютерная лингвистика: научное направление и учебная дисциплина: сборник научных статей. Вып. 1 / В.И. Коваль (ответств. ред.) [и др.]; М-во образования РБ, ГГУ им. Ф. Скорины. – Гомель: ГГУ им. Ф. Скорины, 2010. – 236 с.
25. Котов, Р.Г. Прикладная лингвистика и информационная технология. Р.Г. Котов, А.И. Новиков, Ю.П. Скокан. – М.: Наука, 1987. – 163 с.
26. Кузнецов А.Л. Образовательные электронные издания и ресурсы: методическое пособие / С.Г. Григорье, В.В. Гриншкун. – М.: Дрофа, 2009. – 156 с.
27. Кутузов, А.Б. Компьютерные технилогии в формировании профессиональной компетенции переводчика / А.Б. Кутузов // Языки профессиональной коммуникации: сборник статей Третьей международной научной конференции, т. 2. – Челябинск, 2007. [Электронный ресурс] – URL: http://tc.utmn.ru/files/kutuzov_it.pdf.
28. Лингвистический энциклопедический словарь / Гл. ред. В.Н. Ярцева. – М.: Сов. Энциклопедия, 1990. – 665 с.
29. Лингвистическое моделирование коммуникативных систем: сб. науч. ст. / МГПИИЯ; редкол.: А.В. Зубов (отв. ред.) [и др.]. – Минск, 1983. – 263 с.

30. Лингвостатистика и автоматический анализ текстов: сб. науч. ст. / МГПИИЯ; редкол.: А.В. Зубов (отв. ред.) [и др.]. – Минск, 1973. – 458 с.
31. Лукашевич, Н.В. Тезаурус в задачах информационного поиска / Н.В. Лукашевич. – М., 2010. – 396 с.
32. Максименко, О.И. Формальные методы в современной прикладной лингвистике О.И. Максименко. – М.: МГОУ, 2002. – 256 с.
33. Марчук, Ю.Н. Информационные технологии в лингвистике: компьютерная лингвистика / Ю.Н. Марчук. – parmarium: Acad. Publishing, 2015. – 131 с.
34. Марчук, Ю.Н. Компьютерная лингвистика: учебное пособие / Ю.Н. Марчук. – М.: АСТ: Восток–Запад, 2007. – 317 с.
35. Марчук, Ю.Н. Методы моделирования перевода / Ю.Н. Марчук. – М.: Наука, 1985. – 203 с.
36. Марчук, Ю.Н. Проблемы машинного перевода / Ю.Н. Марчук. – М.: Наука, 1983. – 232 с.
37. Методы анализа текстов: сб. науч. ст. / МГПИИЯ; редкол.: А.В. Зубов (отв. ред.) [и др.]. – Минск, 1975. – 226 с.
38. Мохаммади, Мохаммад Реза. Система русско-персидского машинного перевода на основе переводных соответствий: дисс. канд. филол. наук / Мохаммад Реза Моххамади. – М.: МГУ, 1998. – 132 с.
39. Нелюбин, Л.Л. Компьютерная лингвистика и машинный перевод / Л.Л. Нелюбин. – М.: Наука, 1983. – 241 с.
40. Новейший словарь иностранных слов / [авт.-сост. Е.А. Окунцова]. – М.: Айрис-пресс, 2007. – 512 с.
41. Новое в зарубежной лингвистике: Вып. XXIV. Компьютерная лингвистика: Пер. с англ. / Сост., ред. и вступ, ст. Б.Ю. Городецкого. – М.: Прогресс, 1989. – 432 с.
42. Орёл, М.А. Словарь переводчику – друг, товарищ и Брут / М.А. Орёл // Перевод: информационные технологии. – М.: Всероссийский центр переводов научно-технической литературы и документации, 2009. – С. 79 – 106.
43. Перевод: традиции и современные технологии [Сб. ст.] / М-во пром-сти, науки и технологий Рос. Федерации, Рос. акад. наук, Всерос. центр пер. науч.-техн. лит. и документации; [Отв. ред. д. филол. н., проф. Убин И.И.]. – М.: ВЦП, 2002. – 131 с.
44. Пиотровский, Р.Г. Инженерная лингвистика и теория языка. / Р.Г. Пиотровский. – Л.: Наука, 1979. – 112 с.
45. Пиотровский, Р.Г. Лингвистический автомат и его речемыслительное обоснование / Р.Г. Пиотровский. – Минск: МГЛУ, 1999. – 196 с.
46. Попов, С.А. Информационные технологии в лингвистике / С.А. Попов, Е.В Жукова; М-во образования и науки Российской Федерации, Федеральное гос. бюджетное образовательное учреждение высшего проф. образования «Новгородский гос. ун-т им. Ярослава Мудрого». – Великий Новгород: Новгородский гос. ун-т, 2014. – 235 с.

47. Потапова, Р.К. Новые информационные технологии и лингвистика / Р.К. Потапова. – М.: МГЛУ, 2002. – 575 с.
48. Потапова, Р.К. Речь: коммуникация, информация, кибернетика: Учеб, пособие для студентов вузов, обучающихся по специальностям «Автоматизир. системы обраб. информ. и упр.», «Лингвистика» / Р.К. Потапова. – 3. изд., стер. – М.: УРСС, 2003. – 564 с.
49. Прикладное языкознание Учебник / [С.А. Аверина, И.В. Азарова, Е.Л. Алексеева и др.]; Отв. ред. А.С. Герд; С.-Петерб. гос. ун-т. – СПб.: Изд-во С.-Петерб. ун-та, 1996. – 525 с.
50. Проблемы компьютерной лингвистики: сб. науч. ст. / МГЛУ; редкол.: Р.Г. Пиотровский (отв. ред.) [и др.]. – Минск, 1997. – 178 с.
51. Пропп, В.Я. Исторические корни волшебной сказки / В.Я. Пропп. – Ленинград: Издательство ЛГУ, 1986. – 370 с.
52. Пропп, В.Я. Морфология сказки / В.Я. Пропп. – Ленинград: Academia, 1928. – 152 с.
53. Рождественский, Ю.В. Введение в общую филологию / Ю.В. Рождественский. – М.: Высшая школа, 1979. – 223 с.
54. Рождественский, Ю.В., Введение в прикладную филологию / Ю.В. Рождественский, А.А. Волков, Ю.Н. Марчук. – М.: МГУ, 1988. – 116 с.
55. Соловьёва, А.В. Профессиональный перевод с помощью компьютера / А.В. Соловьёва. – СПб.: Литер, 2008. – 160 с.
56. Соснина Е.П. Введение в прикладную лингвистику / Е.П. Соснина. – Ульяновск, 2010. [Электронный ресурс]. – URL:http://www.twirpx.com/file/736011/ – электронный учебник.
57. Структурализм: «за» и «против2: Сборник статей: Пер. с англ., фр., нем., чеш., польск. и болг. яз. / Под ред. Е.Я. Басина и М.Я. Полякова; [Предисл. В.П. Крутоуса, с. 3–24]. – Москва: Прогресс, 1975. – 468 с.
58. Частные вопросы автоматического анализа текстов: сб. науч. ст. / МГПИИЯ; редкол.: А.В. Зубов (отв. ред.) [и др.]. – Минск, 1972. – 395 с.
59. Частотный словарь русского языка: Около 40000 слов / [Сост. В.А. Аграев, В.В. Бородин, Л.Н. Засорина и др.]; под ред. Л.Н. Засориной. – М.: Рус. яз, 1977.
60. https://bnkorpus.info/
61. https://ruscorpora.ru/
62. https://www.baidu.com/
63. https://www.google.by/
64. https://www.nigma.net.ru/
65. https://www.rambler.ru/
66. https://www.yahoo.com/
67. https://yandex.ru/

ПРАКТИЧЕСКИЙ РАЗДЕЛ

МОДУЛЬ I

Тема 1
ИНФОРМАЦИЯ И ЛИНГВИСТИКА

Вопросы, рассматриваемые на занятии:

1. Прикладная лингвистика как научное направление.
2. Лингвистика и информация.
3. Лингвистика и информатика.
3. Методы прикладной лингвистики.
4. Традиционные и новые задачи прикладной лингвистики.

Задание 1. Подготовьте развернутые ответы на следующие вопросы.

1. Что такое прикладная лингвистика? Каковы особенности понимания этого термина в англоязычной и русскоязычной традициях?

2. В каких аспектах и как изучается информация современными лингвистическими направлениями?

3. Каковы сферы пересечения информатики и лингвистики? Как взаимодействуют эти две дисциплины?

4. Каков методологический аппарат прикладной лингвистики?

Задание 2. Каким терминам и понятиям соответствуют следующие определения? Учтите, что некоторые ответы могут повторяться.

1) Отрасль науки, изучающая общие свойства информации, а также вопросы, связанные с её накоплением, преобразованием, поиском, хранением и передачей с помощью компьютеров и других вычислительных средств.

2) Лингвистическое направление, задачей которого является разработка способов и приемов обучения иностранному языку, включая методику преподавания иностранного языка, особенности описания грамматики в учебных целях и т. п.

3) Направление в языкознании, занимающееся разработкой методов решения практических задач, связанных с использованием языка.

4) Сфера практического применения вычислительной техники.

5) Сообщение о фактах, событиях, о состоянии чего-либо.

6) Наука о естественном человеческом языке вообще и о всех языках мира как индивидуальных его представителях.

7) Материальный объект, предназначенный для хранения данных.

8) Область информатики, занимающаяся вопросами ее лингвистического обеспечения.

9) Область информатики, занимающаяся вопросами ее аппаратного обеспечения.

10) Область информатики, занимающаяся вопросами ее программного обеспечения.

11) Наука, представители которой исследуют язык с помощью точных математических методов и компьютерных программ.

12) Совокупность сведений как объект хранения, переработки и передачи.

Для справок: *информатика*, *информация*, *квантитативная лингвистика*, *лингвистика*, *носитель информации*, *прикладная лингвистика*, *applied linguistics, hardware, lingware, software*.

Задание 3. Подготовьте сообщение на тему «Создание письменностей на основе кириллицы в XX веке», в котором расскажите о практике советского языкознания в этой области. В сообщении отметьте следующие моменты:

1) причины отсутствия письменности у разных народов;

2) эффективность использования кириллического письма для разных языков;

3) судьба письменных систем после распада Советского Союза.

Задание 4. Познакомьтесь с фрагментом текста, рассказывающем об истории создания первой азбуки для слепых, и подготовьте развернутое сообщение о принципах и особенностях такого письма.

«Рельефно-точечный тактильный шрифт, предназначенный для письма и чтения незрячими и плохо видящим людям. Разработан в 1824 году французом Луи Брайлем, сыном сапожника. Луи в возрасте трёх лет поранился в мастерской отца шорным ножом; из-за начавшегося воспаления глаза мальчик потерял зрение. В возрасте 15 лет Луи создал свой рельефно-точечный шрифт как альтернативу рельефно-линейному шрифту Валентина Гаюи, вдохновившись простотой «ночного шрифта» капитана артиллерии Шарля Барбье. В то время «ночной шрифт» использовался военными для записи донесений, которые можно было прочесть в темноте. Шрифт Брайля был первой системой записи с двоичным кодированием».

Задание 5. Подготовьте развернутое сообщение на тему «Проблемы упорядочения, унификация и стандартизации научно-технической терминологии».

Задание 6. Прокомментируйте следующие типы моделей, на которые ориентируется теоретическая лингвистика при использовании метода моделирования:

1) компонентные;

2) предсказывающие;

3) имитирующие;

4) диахронические.

Задание 7. Подготовьте развернутое сообщение на тему:

1) «Лингвистические основы машинного перевода».

2) «Лингвистические основы автоматического анализа и синтеза текстов».

Тема 2
ЛИНГВИСТИЧЕСКИЕ ОСНОВЫ ИНФОРМАТИКИ И КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

Вопросы, рассматриваемые на занятии:

1. Моделирование общения.
2. Моделирование сюжета.
3. Метауровень коммуникации и его роль в моделировании общения.
4. Типы коммуникативного общения и их учет в моделировании коммуникации.
5. Синтаксис сюжета.
6. Когнитивный подход к моделированию.

Задание 1. Подготовьте развернутые ответы на следующие вопросы:

1. Метод моделирования в компьютерной лингвистике.
2. Способы моделирования общения: история и современные подходы.
3. Уровни коммуникации. Уровень метакоммуникации.
4. Построение синтаксиса сюжета.
5. Сущность когнитивного подхода в моделировании общения и сюжета.

Задание 2. Каким терминам и понятиям соответствуют следующие определения?

1) Научное направление, которое рассматривает язык как знаковую систему с четко выраженными структурными элементами.

2) Вторичная коммуникация, включающая косвенные сигналы о том, как предполагается интерпретировать фрагмент инфоррмации, коммуникация «по поводу коммуникации».

3) Исследование каких-либо объектов на их моделях; построение моделей реально существующих предметов, явлений или процессов (живых организмов, инженерных конструкций, образцов одежды и т.п.).

4) Приемопередающий буквопечатающий телеграфный аппарат, который применяется для передачи сообщений по каналам связи, а также в качестве терминала в устройствах вычислительной техники.

5) Направление в науке, задачей которого является обнаружение, описание и объяснение структуры мышления, лежащей в основе культуры прошлого и настоящего.

6) Направление компьютерной лингвистики, задачей которого является представление в формальном виде общение человека с персональным компьютером.

7) Формализм генеративной лингвистики, связанный с изучением синтаксиса.

8) Устройство для ввода и вывода информации (клавиатура, дисплей, принтер и др.).

9) Самостоятельно созданное повествование о некотором множестве взаимосвязанных событий, представленное читателю или слушателю в виде последовательности слов или образов.

10) Операция замены смысловых содержательных описаний некоторых явлений, процессов, состояний математическими структурами.

Для справок: *метакоммуникация*, *моделирование*, *моделирование общения*, *нарратив*, *порождающая грамматика*, *структурная лингвистика*, *структурное литературоведение*, *телетайп*, *терминал*, *формализм*.

Задание 3. Прочтите фрагмент словарной статьи и вставьте на месте пропусков подходящие по смыслу термины. Прокомментируйте встретившиеся в тексте понятия «когнитивная грамматика», «функционалистская теория», «бихевиористская теория». Расскажите об истории возникновения и дальнейшей судьбе этого научного направления.

«В рамках <…?> формулируется система правил, при помощи которых можно определить, какая комбинация слов оформляет грамматически правильное предложение. Термин введен в научный оборот в работах Ноама Хомского в конце 1950-х годов. Хомский утверждает, что многие свойства <…?> производны от универсальной грамматики. Этим <…?> отличается от подходов, принятых в когнитивной грамматике, функционалистской и бихевиористской теориях».

Задание 4. Прочтите фрагмент статьи из «Лингвистического энциклопедического словаря», посвященной понятию «Генеративная лингвистика», и выполните размещенные после него задания.

«Трансформационная порождающая грамматика описывает прежде всего компетенцию говорящего. Структура этой грамматики имеет три основных компонента: синтаксический, семантический и фонологический, из которых главным, центральным является <…?>, а <…?> и <…?> выполняют по отношению к <…?> интерпретирующие функции» [28, с. 98].

1) Дайте краткую характеристику генеративной лингвистике как научному направлению.

2) Вставьте на месте пропусков термины ʻсемантикаʼ, ʻсинтаксисʼ, ʻфонологияʼв соответствии с логикой изложения. Аргументируйте свой выбор, опираясь на идеи генеративной лингвистики.

Задание 5. Прочтите фрагмент работы В.Я. Проппа «Морфология волшебной сказки» и выполните задания (ответьте на вопросы), размещенные после него.

«В 1924 г. появилась книга о сказке одесского профессора Р.М. Волкова. Волков с первых же страниц своего труда определяет, что фантастическая сказка знает 15 сюжетов. Сюжеты эти следующие:

1) О невинно гонимых.
2) О герое-дурне.
3) О трех братьях.
4) О змееборцах.
5) О добывании невест.
6) О мудрой деве.
7) О заклятых и зачарованных.
8) Об обладателе талисмана.
9) Об обладателе чудесных предметов.
10) О неверной жене и т.д.

Как установлены эти 15 сюжетов – не оговорено. Если же всмотреться в принцип деления, то получится следующее: первый разряд определен по завязке (что здесь действительно завязка, мы увидим ниже), второй — по характеру героя, третий – по количеству героев, четвертый – по одному из моментов хода действия и т.д. Таким образом, принцип деления вообще отсутствует. Получается действительно хаос. Разве нет сказок, где три брата (третий разряд) добывают себе невест (пятый разряд)? Разве обладатель талисмана не наказывает с помощью этого талисмана неверную жену? Таким образом, данная классификация не является научной классификацией в точном смысле слова, она не более как условный указатель, ценность которого весьма сомнительна. И разве может подобная классификация хотя бы отдаленно сравниваться с классификацией растений или животных, произведенной не на глаз, а после точного и длительного предварительного изучения материала?»

1) Приведите по два-три примера сказок, построенных по каждому из перечисленных во фрагменте сюжетов.

2) Согласны ли вы с мнением о том, что количество сказочных сюжетов весьма ограниченно?

3) Согласны ли вы с оценкой классификации сказок Р.М. Волкова, данной В.Я. Проппом? Аргументируйте свою точку зрения.

Задание 6. Перечислите выделенные В.Я. Проппом постоянные элементы сказки. Проиллюстрируйте ответ примерами сказок, в которых данные элементы нашли реализацию. Назовите компьютерную программу, основу которой составили идеи В.Я. Проппа. Почему чисто морфологический подход к созданию подобных программ является недостаточным? Найдите дополнительную информацию о моделировании сюжета современными компьютерными программами.

Тема 3
КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

Вопросы, рассматриваемые на занятии:

1. Компьютерная лингвистика как современное научное направление.
2. Межпредметные связи компьютерной лингвистики.
3. Теоретическое направление компьютерной лингвистики.
4. Прикладное направление компьютерной лингвистики и его задачи.

Задание 1. Подготовьте развернутые ответы на следующие вопросы:

1. Что такое компьютерная лингвистика? Как соотносятся между собой понятия ʻкомпьютерная лингвистикаʼ, ʻинформационная лингвистикаʼ, ʻприкладная лингвистикаʼ?

2. Каковы межпредметные связи компьютерной лингвистики? В чем именно состоят зоны пересечения компьютерной лингвистики с каждым из направлений?

3. Каковы задачи теоретического направления компьютерной лингвистики?

4. В чем состоят цели и перспективы прикладного направления компьютерной лингвистики?

Задание 2. Каким терминам и понятиям соответствуют следующие определения?

1) Мыслительная способность человека, уровень умственного развития, способность к рациональному познанию.

2) Отрасль науки, изучающая общие свойства информации, а также вопросы, связанные с её накоплением, преобразованием, поиском, хранением и передачей с помощью компьютеров и других вычислительных средств.

3) Наука о знаниях, приемах и методах получения, обработки, хранения и использования знаний.

4) Наука о математических методах и алгоритмах преобразования данных для обеспечения их конфиденциальности, целостности и аутентификации.

5) Область сравнительно-исторического языкознания, занимающаяся выявлением скорости языковых изменений и определением на этом основании времени разделения родственных языков и степени близости между ними.

6) Всемирная информационная компьютерная сеть, связывающая между собой как пользователей компьютерных сетей, так и пользователей индивидуальных компьютеров, позволяющая обмениваться разнообразной информацией.

7) Совокупность сведений как объект хранения, переработки и передачи.

8) Один из методов сравнительной лингвистики, применяемый для сравнения процентного соотношения родственных слов разных языков и определения взаимосвязи между словами и языками.

9) Направление в лингвистике, которое объединяет исследования в области компьютерной лингвистики и машинного обучения.

10) Наука, исследующая язык с помощью точных математических методов и компьютерных программ.

Для справок: *глоттохронология*, *интеллект*, *интернет*, *информатика*, *информация*, *искусственный интеллект*, *квантитативная лингвистика когнитология*, *криптография лексикостатика*, *логика*.

Задание 3. Подготовьте развернутое сообщение на тему «Искусственный интеллект и его использование в научных исследованиях гуманитарного профиля».

Задание 4. Перечислите современные электронные словари, тезаурусы, онтологии. Изучите устройство одного из таких ресурсов и расскажите о нем.

Задание 5. Назовите известные вам программы-переводчики. Изучите детальнее принципы и особенности их работы. Расскажите о собственном опыте использования программ автоматизированного перевода.

Задание 6. Переведите на белорусский или другой, используя одну из автоматизированных программ, следующие фрагменты текстов разных жанров и типов. Дайте лингвистическую оценку качеству автоматизированного перевода.

1) «*Документирование управленческой деятельности заключается в создании управленческих документов – фиксации на бумаге или других носителях управленческих действий по установленным правилам. Документирование управленческой деятельности может осуществляться как рукописным способом, так и с помощью технических средств.*

Состав документов, образующихся в деятельности организации, определяется ее компетенцией, кругом управленческих функций, порядком

разрешения вопросов (единоличный или коллегиальный), объемом и характером взаимосвязей между организациями одного или различных уровней управления и т.п.

Юридическим основанием создания организационно-распорядительных документов в деятельности организаций являются:

акты законодательства Республики Беларусь;

решения судов;

предписания государственных органов и должностных лиц;

поручения вышестоящих организаций;

осуществление исполнительной и организационно-распорядительной деятельности в целях выполнения организацией возложенных на нее функций и задач в соответствии с ее компетенцией».

2) «*Все знают, как важно беречь в облике нашей земли, нашей страны всё то, что может тронуть человеческое сердце и оставить о себе благодарную память. Однако, если присмотреться, мы всё-таки плохо бережём то, что непременно надо беречь. Например, частенько на «клочке земли, припавшем к трём берёзам», можно увидеть кучу бетонного мусора, или забытую ржавую сеялку, или «нетленную» кучу полиэтиленовых мешков из-под удобрений. Опушки лесов и полосы лесопосадок в степи опалены химикатами, неаккуратно распылёнными с самолёта. Деревья в лесу испачканы пятнами масляной краски – помечали туристический маршрут.*

Или ещё такой пример: зелёный травяной склон горы нередко изрезан громадными буквами какого-нибудь призыва, например: «Берегите лес!». Достигает ли цели этот призыв, когда его вырубают лопатой в зелёном дёрне? После всего этого даже как-то ***неловко*** *говорить о тонкостях восприятия человеком красоты пейзажа*».

3) «*За рекой синел лес. Репин знал, что этот цвет синевы обманчив. Выглянет из-за облачка солнце – и первый же луч все изменит. А он бессилен передать на полотне это движение, смену цветов и оттенков в природе.*

Художник еще раз коснулся кистью полотна, обвел кромку леса и бросил чуть-чуть желтизны не деревья. Прищурился, поглядел…

У ног его на августовской росе вспыхнуло солнце, потом рассыпалось на искринки и весело заиграло в желтых листьях. И, словно от его прикосновения, листья зашуршали. Репин сделал несколько быстрых движений кистью. Затем рука замедлила полет. Внимание художника отвлек падающий лист» (Из книги Д. Симановича «Сквозь даль времен»).

4) *«Во многом аналогична ситуация и с употреблением форм существительных. В обоих языках, русском и английском, информация о количестве объектов входит в «грамматическую анкету» (хотя правила употребления граммем единственного и множественного числа в некоторых тонких деталях различаются – здесь еще один источник расхождения между грамматическими системами разных языков). Но в английском языке при употреблении любого существительного, кроме этого,*

дополнительно требуется ответить и на вопрос о его «детерминации»: каждое английское существительное обязательно сопровождается в тексте либо определенным, либо неопределенным артиклем (либо не сопровождается никаким, но это отсутствие артикля в данном случае тоже имеет строго определенную функцию)» (В.А. Плунгян «Введение в грамматическую семантику»).

Задание 7. Переведите на русский язык фрагмент романа В. Короткевича «Черный замок Ольшанский» на русский, используя автоматизированную программу. Сравните полученный перевод с реальным, выполненным профессиональным переводчиком.

«Прозвішча маё Касміч. Хрышчаны (гэта ўсё бабця) Антонам. Бацьку, калі хочаце ведаць, звалі Глебам. Маці – Багуславай. Занятак бацькоў да рэвалюцыі? Бацька, скончыўшы гімназію, якраз паспеў на Зялёнага і Махно. Маці – гады тры-як кончыла гуляць у лялькі.

Я даю вам гэтую разгорнутую анкету, каб не гаманіць доўга. Анкету з дабаўленнем міліцэйскага апісання прыкмет. А раптам нешта нараблю. Асабліва пры маёй схільнасці ўблытвацца ў розныя прыгоды, на якія мне да таго ж шанцуе.

Мне трыццаць восем год без малога. Стары кавалер, як казала мая знаёмая Зоя Пярэрвенка, з якой у мяне тады якраз канчаўся – а мабыць, такі канчаўся – кароткі і, як заўсёды, не вельмі ўдалы раман. Гэткі стары-ы кавалер, які за вайной ды навукай не ажаніўся, а цяперака позна ўжо».

«Фамилия моя Космич. Крещенный (это все бабка) Антоном. Отца, если хотите, звали Глебом. Мать – Богуславой. Занятие родителей до революции? Отец, окончив гимназию, как раз успел на Зеленого и Махно. Мать – года три как перестала играть в куклы. Я даю вам эту развернутую анкету для того, чтобы все было ясно. Анкету с добавлением милицейского описания примет. А вдруг чего-то натворю? Особенно при моей склонности впутываться в различные приключения, на которые мне к тому же везет. Мне без малого тридцать восемь лет. Старый кавалер, как говорила моя знакомая Зоя Перервенко, с которой у меня тогда как раз кончался короткий и, как всегда, не очень удачный роман. Этакий старый холостяк, который из-за войны да науки не женился, а теперь уже поздно».

Задание 8. Переведите на белорусский или другой, используя одну из автоматизированных программ, фрагменты «Российской грамматики» М.В. Ломоносова. Дайте лингвистическую оценку качеству автоматизированного перевода. Сравните качество перевода современных текстов (задание 6) с качеством перевода текста, написанного на старом языке.

«*По благороднейшем даровании, которым человек прочих животных превосходит, то есть правителе наших действий — разуме, первейшее есть слово, данное ему для сообщения с другими своих мыслей. Польза его*

толь велика, коль далече ныне простираются происшедшие от него в обществе человеческом знания, которые весьма бы тесно ограничены были, если бы каждый человек воображенные себе способом чувств понятия только в собственном своем уме содержал сокровенны. Когда к сооружению какой-либо махины приготовленные части лежат особливо, и никотора я определенного себе действия другой взаимно не сообщает, тогда все бытие их тщетно и бесполезно. Подобным образом, если бы каждый член человеческого рода не мог изъяснить своих понятий другому, то бы не токмо лишены мы были сего согласного общих дел течения, которое соединением разных мыслей управляется, но и едва бы не хуже ли были мы диких зверей, рассыпанных по лесам и по пустыням».

МОДУЛЬ II

Тема 1
КОРПУСНАЯ ЛИНГВИСТИКА

Вопросы, рассматриваемые на занятии:

1. Лингвистический корпус.
2. Корпусная лингвистика как современное научное направление.
3. Задачи корпусной лингвистики.
4. Современные корпусы национальных языков.

Задание 1. Подготовьте развернутые ответы на следующие вопросы:

1. Что такое языковой корпус? Чем он отличается от других способов представления собраний текстов в Интернете?

2. Какие цели ставят создатели современных корпусов? Какие функции они способны выполнять?

3. В чем специфика представления языка в Национальном корпусе русского языка?

4. На какой стадии находится разработка корпуса белорусского языка?

Задание 2. Каким терминам и понятиям соответствуют следующие определения?

1) Речь, связный текст в совокупности с прагматическими, социокультурными и другими факторами; речь как целенаправленное социальное действие.

2) Программа, предоставляющая удобный доступ к базе данных корпуса.

3) Тип словаря, в котором все слова данного языка представлены максимально полно и исчерпывающим перечнем их употребления в тексте.

4) Специальный раздел языкознания, который занимается проблемами разработки, создания и использования текстовых корпусов.

5) Тип словаря, в котором все слова данного языка представлены максимально полно и исчерпывающим перечнем их употребления в тексте.

6) Большой, представленный в машиночитаемом формате, унифицированный, структурированный, размеченный филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач.

Для справок: *дискурс*, *корпусная лингвистика*, *корпусный менеджер*, *тезаурус*, *языковой корпус*.

Задание 3. Изучите возможности, предоставляемые белорусским N-корпусом, и выполните следующие задания.

1) Выберите все имена существительные из одного-двух литературных произведений. Задайте в системе поиска определение их грамматических особенностей.

2) Проведите исследование новых слов, недавно появившихся в белорусском языке и еще не отраженных в нормативных словарях.

3) Введите самостоятельно созданный текст для проверки его орфографии.

Задание 4. Предложите идеи для усовершенствования возможностей белорусского N-корпуса. Сформулируйте темы возможных исследований, направленных на решение этой задачи.

Задание 5. Изучите возможности, предоставляемые Национальным корпусом русского языка. Проведите лингвистическое исследования, связанное с вашими научными интересами. Дайте оценку эффективности работы с корпусом.

Тема 2
СОВРЕМЕННЫЕ ПОИСКОВЫЕ СИСТЕМЫ

Вопросы, рассматриваемые на занятии:

1. Поиск информации в Интернете.
2. Современные поисковые системы.
3. Информационно-поисковые тезаурусы.
4. Лингвистические дескрипторы и их особенности.

Задание 1. Подготовьте развернутые ответы на следующие вопросы.

1. Что такое информационный поиск?
2. Что такое поисковая система и поисковая машина?

3. Какие этапы включает в себя информационный поиск? В чем специфика каждого из этапов?

4. В чем основные различия между поиском в сети Интернет и обычном информационным поиском?

5. Какие факторы учитывает поисковая машина при отборе запрашиваемой пользователем информации?

Задание 2. Каким терминам и понятиям соответствуют следующие определения?

1) Автоматизированный программно-аппаратный комплекс с веб-интерфейсом, предоставляющего возможность поиска информации в Интернете.

2) Система унифицированных связей и сигналов, посредством которых устройства вычислительной системы взаимодействуют друг с другом.

3) Словарь, в котором слова и словосочетания с близкими значениями сгруппированы в единицы, называемые понятиями, концептами или дескрипторами, и в котором явно (в виде отношений, иерархии) указываются семантические отношения между этими понятиями (концептами, дескрипторами.

4) Лексическая единица (слово, словосочетание) информационно-поискового языка, служащая для описания основного смыслового содержания документа или формулировки запроса при поиске документа (информации) в информационно-поисковой системе.

5) Лексическая единица в информационно-поисковом тезаурусе, которая не может быть использована для координатного индексирования и подлежит замене одним или несколькими дескрипторами.

6) Специальное программное обеспечение для организации общения посетителей сайта.

7) Программа для поиска и просмотра на экране компьютера информации из компьютерной сети; программа просмотра; навигатор.

8) Символ или слово, которые используются для различения значений многозначного термина.

9) Отдельный документ в Интернете, который может содержать текст, графику, звук и т.д.

10) Страница в Интернете, оформленная в виде журнала, дневника, основное содержание которого – постоянно добавляемые записи, изображения и мультимедиа.

Для справок: *аскриптор, браузер, блог, веб-интерфейс, веб-страница, веб-форум, дескриптор, поисковая система, релятор, тезаурус.*

Задание 3. Ознакомьтесь с правилами формирования запросов в поисковой системе Яндекс, сформулированными А. В. Гусляковой. Используя их, организуйте поиск информации по теме вашего исследования. Дайте оценку эффективности поиска, осуществленного по данным правилам.

«1. Ключевые слова в запросе следует писать строчными (маленькими) буквами. Это обеспечит поиск всех ключевых слов, а не только тех, которые начинаются с прописной буквы.

2. При поиске учитываются все формы слова по правилам русского языка, независимо от формы слова в запросе. Например, если в запросе было указано слово «знаю», то условию поиска будут удовлетворять и слова «знаем», «знаете» и т. и.

3. Для поиска устойчивого словосочетания следует заключить слова в кавычки. Например, «фонема».

4. Для поиска по точной словоформе перед словом надо поставить восклицательный знак. Например, для поиска слова «сентябрь» в родительном падеже следует написать «¡сентября».

5. Для поиска внутри одного предложения слова в запросе разделяют пробелом или знаком &. Например, «приключенческий роман» или «приключенческий&роман». Несколько набранных в запросе слов, разделенных пробелами, означают, что все они должны входить в одно предложение искомого документа.

6. Для того, чтобы были отобраны только те документы, в которых встретилось каждое слово, указанное в запросе, необходимо поставить перед каждым из них знак плюс «+». Если вы, наоборот, хотите исключить какие-либо слова из результата поиска, поставьте перед этим словом минус «-». Знаки «+» и «-» надо писать через пробел от предыдущего и слитно со следующим словом. Например, по запросу «Волга – автомобиль» будут найдены документы, в которых есть слово «Волга» и нет слова «автомобиль».

7. При поиске синонимов или близких по значению слов между словами можно поставить вертикальную черту «|». Например, по запросу «ребенок | малыш | младенец» будут найдены документы с любым из этих слов.

8. Вместо одного слова в запросе можно подставить целое выражение. Для этого его надо взять в скобки. Например, «(ребенок | малыш | дети | младенец) +(уход | воспитание)».

9. Знак «~» (тильда) позволяет найти документы с предложением, содержащим первое слово, но не содержащим второе. Например, по запросу «книги – магазин» будут найдены все документы, содержащие слово «книги», рядом с которым (в пределах предложения) нет слова «магазин».

10. Если оператор повторяется один раз (например, & или ~), поиск производится в пределах предложения. Двойной оператор (&&,—) задает поиск в пределах документа. Например, по запросу «дева – астрология» будут найдены документы со словом «дева», не относящиеся к астрологии.

11. Вернемся к примеру с аквариумными рыбками. После прочтения нескольких предлагаемых поисковой системой документов становится понятно, что поиск информации в Интернете следует начинать не с выбора аквариумных рыбок. Аквариум – сложная биологическая система, создание и поддержание которой требует специальных знаний, времени и серьезных капиталовложений» [12, с. 68].

Задание 4. Организуйте поиск необходимой для проведения вашего исследования информации в разных поисковых системах. В случае обнаружения отличий в результатах поиска охарактеризуйте их.

Задание 5. Используя две-три различные поисковые системы, составьте список библиографических источников по теме вашего исследования. Сравните результаты поиска.

Задание 6. Ознакомьтесь с компонентами поисковой машины. Найдите в Интернете дополнительную информацию об их работе.

1. Паук или спайдер (spider). Приложение, которое занимается скачиванием страниц Интернет-ресурсов. «Паук» запрашивает содержимое страниц точно так же, как это делает обычный интернет-браузер, отправляя на сервер HTTP запрос и получая от него ответ. После того, как содержимое страницы скачано, оно отправляется индексатору и краулеру, о которых рассказывается далее.

2. *Индексатор (indexer).* Индексатор производит первоначальный анализ содержимого скачанной страницы, выделяет основные части (название страницы, описание, ссылки, заголовки и т.д.) и раскладывает все это по разделам поисковой базы данных – помещает в индекс поисковой системы. Этот процесс называют индексацией интернет ресурсов, отсюда и название самой подсистемы. На основе результатов первоначального анализа индексатор также может принять решение, что страница вообще «недостойна» находиться в индексе. Причины такого решение могут быть разными: страница не имеет названия, является точной копией другой, уже имеющейся в индексе страницы или содержит ссылки на запрещенные законодательством ресурсы.

3. *Краулер (crawler).* Это приложение призвано перемещаться по ссылкам, имеющимся на скачанной пауком странице. Краулер анализирует пути, ведущие с текущей страницы на другие разделы сайта, или на страницы внешних Интернет ресурсов и определяет дальнейший порядок обхода пауком нитей всемирной паутины. Именно краулер находит новые для поисковой машины страницы и передает их пауку. Работа краулера построена на базе алгоритмов поиска на графах в ширину и глубину.

4. *Подсистема обработки и выдачи результатов (Search Engine and Results Engine).* Самая важная часть любой поисковой машины. Алгоритмы работы этой подсистемы компании разработчики хранят в строгой секретности, поскольку они являют собой коммерческую тайну. Именно эта часть поисковой машины отвечает за адекватность ответа поисковой системы на запрос пользователя. Здесь можно выделить два основных компонента:

• *Подсистема ранжировапия.* Ранжирование – это сортировка страниц интернет сайтов в соответствии с их релевантностью определенному запросу. Релевантность страницы – это, в свою очередь, степень

соответствия содержания страницы смыслу запроса, и эту величину поисковая машина определяет самостоятельно, исходя из огромного количества параметров. Ранжирование – эта самая загадочная и спорная часть «искусственного интеллекта» поисковой машины. На ранжирование страницы, помимо ее структуры и содержимого (контента) также влияют: количество и качество ссылок, ведущих на данную страницу с других сайтов; возраст домена самого сайта; характер поведения пользователей, просматривающих страницу и многие другие факторы.

• *Подсистема выдачи результатов.* В задачи этой подсистемы входит интерпретация пользовательского запроса, его перевод на язык структурированных запросов к индексу и формирование страниц результатов поиска.

Помимо разбора самого текста запроса, поисковая машина может также учитывать *контекст запроса*, формируемый исходя из смысла ранее осуществленных пользователем запросов. Так, например, если пользователь часто посещает сайты на автомобильные темы, то на запрос со словом «Волга» или «Ока» он, вероятно, хочет получить информацию об автомобилях этих марок, а не о том, откуда начинают свое течение и куда впадают одноименные русские реки. Это называется персонализированным поиском, когда выдача на один и тот же запрос для разных пользователей существенно отличается» [12, 2016, с. 29].

РАЗДЕЛ КОНТРОЛЯ ЗНАНИЙ

ВОПРОСЫ К ЗАЧЕТУ

1. Новые реалии коммуникации. Современное понимание информации.
2. Причины и факторы возникновения информационной лингвистики. Область возникновения лингвистических проблем информатики.
3. Место информационной лингвистики среди других дисциплин. Межпредметные связи информационной лингвистики. Информационная лингвистика и современное языкознание.
4. Теоретическая и прикладная лингвистика. Разное понимание и разный объем содержания понятия ‘прикладная лингвистика’ в различных подходах и традициях.
5. Задачи современной прикладной лингвистики.
6. Лингвистические основы информатики и компьютерная лингвистика.
7. Теория искусственного интеллекта.
8. Моделирование общения в информационной лингвистике.
9. Моделирование структуры сюжета.
10. Языковой корпус и корпусная лингвистика.
11. Современные корпусы национальных языков.
12. Корпус белорусского языка.
13. Национальный корпус русского языка.
14. Совершенствование массовой и индивидуальной коммуникации.
15. Компьютерная лингвистика и ее задачи. Теоретические и прикладные аспекты компьютерной лингвистики.
16. Информатика и компьютерная лингвистика.
17. Лингвистические аспекты распознавания устной речи компьютерными программами.
18. Автоматическое распознавание звуков устной речи. Проблема распознавания изолированных слов.
19. Автоматический анализ и синтез устной речи.
20. Лингвистические аспекты распознавания письменной речи. Распознавание графем. Исправление искаженных знаков текста.
21. Автоматическая обработка вербального текста.
22. собенности естественного вербального текста. Автоматическое распознавание текста.
23. Диагностика искажений в словах. Автоматический анализ и синтез текста.

24. Лингвистическая дешифровка как прикладная дисциплина. Графематический уровень работы с текстом. Дериватология и ее роль в дешифровке текста.
25. Статистические методы компьютерной лингвистики.
26. Технологии обработки естественного языка в науке и промышленности. Ввод речи (текста) в компьютер.
27. Обработка лингвистической информации на уровне словоформ, слов, словосочетаний, предложений, текста. Машинная морфология.
28. Автоматический морфологический анализ и его виды. Проблемы слова.
29. Вычислительная лексикография. Традиционная и машинная лексикография. Отличия между традиционным и машинным типами словарей.
30. Словарноцентрический подход в компьютерной лингвистике. Автоматический синтаксический анализ.
31. Лемматизация. Машиночитаемые словари.
32. Машинный перевод как аспект современной прикладной компьютерной лингвистики и центральная проблема искусственного интеллекта.
33. Машинный перевод и теория языка. История развития машинного перевода.
34. Современное состояние машинного перевода. Современные автоматизированные программы перевода.
35. Проблемы и перспективы развития автоматического перевода. Преодоление языковых барьеров в машинном переводе. Общая стратегия разработки систем машинного перевода.
36. Компьютерная лингвистика как база для обучения языкам. Современные обучающие программы. Дидактические аспекты компьютерной лингвистики.
37. Специфика использования компьютерных программ для обучения языкам на занятиях разного типа.
38. Современные информационно-поисковые системы. Экспертные системы.
39. Компьтерно-опосредованная коммуникация: феноменологический аспект. Информация компьютерно-опосредованной коммуникации как метаязыковой феномен и коммуникативный процесс, протекающий в открытом электронном социальном окружении.
40. Текст, графика, аудио- и видеофайлы, рисунки и др. как средства компьютерно-опосредованной коммуникации. Специфика имен собственных в компьютерно-опосредованной коммуникации.
41. Информация и дискурс. Компьютерный дискурс.
42. Метаязыковые модели речевой практики. Социокультурная специфика современного интернет-дискурса.
43. Интернет-язык и его социолекты.
44. Лексический и фразеологический корпусы интернет-языка.
45. Проблемы интерпретации и изучения компьютерно-опосредованной коммуникации.

46. Идентичность современной разговорной практики. Номинативная динамика. Деривационная динамика.
47. Репрезентация коммуникации. Лингвоинформационная специфика компьютерно-опосредованной коммуникации. Понятие лингвоинформативности.
48. Тенденции развития разговорной среды. Языковая глобализация. Демократизация речи. Активная неологизация и развитие у слов новых значений.
49. Типология сайтов. Профессиональные сайты. Сайты, ориентированные на филологов.
50. Современные лингвистические исследование, построенные на интернет-материале: проблематика, тематика, методология.

КОНТРОЛЬНЫЙ ТЕСТ

1. Наука о естественном человеческом языке вообще и о всех языках мира как индивидуальных его представителях – это:
   1) теоретическая лингвистика;
   2) прикладная лингвистика;
   3) традиционная лингвистика;
   4) компьютерная лингвистика;
   5) когнитивная лингвистика.

2. Термином *lingware* в информатике обозначают:
   1) аппаратное обеспечение;
   2) программное обеспечение;
   3) лингвистическое обеспечение;
   4) один из языков программирования;
   5) коммуникацию между человеком и компьютером.

3. Ответ на вопрос «Каков язык?» должна дать:
   1) теоретическая лингвистика;
   2) прикладная лингвистика;
   3) компьютерная лингвистика;
   4) когнитивная лингвистика;
   5) социолингвистика.

4. Ответ на вопрос «Как лучше использовать язык?» должна дать:
   1) теоретическая лингвистика;
   2) прикладная лингвистика;
   3) компьютерная лингвистика;
   4) когнитивная лингвистика;
   5) социолингвистика.

5. Сформировавшееся в 1930–40 гг. направление, получившее английское название ‘applied linguistics’, в основном сориентировано на:

1) моделирование общения;
2) моделирование сюжетов;
3) создание искусственных языков;
4) разработку языков программирования;
5) представление языкового материала в учебно-методических целях для изучения иностранных языков.

6. Отрасль науки, изучающая общие свойства информации, а также вопросы, связанные с её накоплением, преобразованием, поиском, хранением и передачей с помощью компьютеров и других вычислительных средств, – это:

1) теоретическая лингвистика;
2) прикладная лингвистика;
3) компьютерная лингвистика;
4) информационная лингвистика;
5) информатика.

7. Любой материальный объект, предназначенный для хранения данных, – это:

1) ЭВМ;
2) компьютер;
3) носитель информации;
4) жесткий диск компьютера;
5) книга.

8. К числу традиционных задач прикладной лингвистики не относится:

1) создание и совершенствование письменностей;
2) создание систем транскрипции устной речи;
3) создание систем транслитерации иноязычных слов;
4) создание систем обучения каллиграфии;
5) создание систем письма для слепых.

9. Методы теории программирования и представления знаний широко используются в:

1) фундаментальной лингвистике;
2) компьютерной лингвистике;
3) когнитивной лингвистике;
4) психолингвистике;
5) социолингвистике.

10. Область сравнительно-исторического языкознания, занимающаяся выявлением скорости языковых изменений и определением на этом основании времени разделения родственных языков и степени близости между ними, – это:

1) глоттохронология;
2) компаративистика;
3) лексикостатика;
4) историческая грамматика;
5) историческая лексикология.

11. Система унифицированных связей и сигналов, посредством которых устройства вычислительной системы взаимодействуют друг с другом, – это:

1) веб-страница;
2) веб-интерфейс;
3) веб-райтер;
4) веб-узел;
5) веб-форум.

12. Литератор, создающий произведения во всемирной компьютерной сети, – это:

1) веб-мастер;
2) веб-ридер;
3) веб-райтер;
4) веб-поэт;
5) веб-прозаик.

13. Автором работы «Синтаксические структуры», которая нашла широкое применение в компьютерной лингвистике, является:

1) Р. Якобсон;
2) В. Я. Пропп;
3) А. Тьюринг;
4) Н. Хомски;
5) К. Леви-Стросс.

14. Первоначальное значение и форма слова, от которого произошло слово современного языка – это:

1) синсет;
2) этимон;
3) тезурус;
4) релятор;
5) лексикостатика.

15. Тип словаря, в котором все слова данного языка представлены максимально полно и исчерпывающим перечнем их употребления в тексте, – это:

1) синсет;
2) этимон;

3) тезурус;
4) релятор;
5) лексикостатика.

16. Один из ключевых элементов в лексической базе данных WordNet и подобных ей системах – это:
1) синсет;
2) этимон;
3) тезурус;
4) релятор;
5) лексикостатика.

17. Символ или слово, которые используются для различения значений многозначного термина, – это:
1) синсет;
2) этимон;
3) тезурус;
4) релятор;
5) лексикостатика.

18. Страница в Интернете, оформленная в виде журнала, дневника, основное содержание которого – постоянно добавляемые записи, изображения и мультимедиа, – это:
1) веб-страница;
2) блог;
3) браузер;
4) веб-интерфейс;
5) тезаурус.

19. Беспроводная персональная сеть, служащая для обмена информацией между цифровыми устройствами на радиочастоте для ближней связи – это:
1) блютуз;
2) блог;
3) браузер;
4) веб-интерфейс;
5) поисковая система.

20. Программа для поиска и просмотра на экране компьютера информации из компьютерной сети; программа просмотра; навигатор – это:
1) блютуз;
2) блог;
3) браузер;
4) веб-интерфейс;
5) поисковая система.

ВСПОМОГАТЕЛЬНЫЙ РАЗДЕЛ

ГЛОССАРИЙ

Аватар – 1) в компьютерных играх – образ, который пользователь выбирает для участия в игре; 2) изображение, картинка, используемая для персонализации пользователя сетевого сервиса.

Аккаунт (эккаунт) – учетная запись, которая заводится на пользователя при регистрации в электронной системе.

Апгрейд – модернизация, усовершенствование чего-либо, замена старого на новое, с лучшими характеристиками.

Байт – в вычислительной технике – единица количества информации или памяти, равная 8 битам.

Блог – страница в Интернете, оформленная в виде журнала, дневника, основное содержание которого – постоянно добавляемые записи, изображения и мультимедиа.

Блютуз – беспроводная персональная сеть, служащая для обмена информацией между цифровыми устройствами на радиочастоте для ближней связи.

Бот – программа, незаконно устанавливаемая на чужом компьютере, позволяющая злоумышленнику выполнять некие действия с использованием зараженного компьютера.

Браузер – программа для поиска и просмотра на экране компьютера информации из компьютерной сети; программа просмотра; навигатор.

Вебинар – лекция, презентация, семинар, проводимые в режиме видеоконференции, во время которой участники слушают выступающего и могут задавать ему вопросы и получать ответы в режиме реального времени.

Веб-интерфейс – система унифицированных связей и сигналов, посредством которых устройства вычислительной системы взаимодействуют друг с другом.

Веб-мастер – программа, которая руководит пользователем при выполнении определенной операции.

Веб-райтер – литератор, создающий произведения во всемирной компьютерной сети.

Веб-ридер – читатель художественного произведения, сообщающий на сайт автора свои замечания.

Веб-сёрфинг – путешествие по компьютерным сетям в поисках нужной или интересной информации.

Веб-страница – отдельный документ в Интернете, который может содержать текст, графику, звук и т.д.

Веб-узел – группа тематически связанных веб-страниц.

Веб-форум – специальное программное обеспечение для организации общения посетителей сайта.

Видеокарта – техническое устройство передачи изображения на монитор.

Виртуальный – 1) возможный, такой, который может или должен проявиться при определенных условиях; 2) воображаемый, реально не существующий, реализованный только в компьютере.

Вторая реальность – воссоздание реальной обстановки с помощью компьютерных устройств.

Глоттохронология – область сравнительно-исторического языкознания, занимающаяся выявлением скорости языковых изменений и определением на этом основании времени разделения родственных языков и степени близости между ними.

Гомо рэтис – человек, посвящающий много времени общению по сетям Интернета.

Дамп – распечатка содержимого памяти ЭВМ или файла, обычно без учета внутренней структуры.

Дескриптор – лексическая единица (слово, словосочетание) информационно-поискового языка, служащая для описания основного смыслового содержания документа или формулировки запроса при поиске документа (информации) в информационно-поисковой системе.

Дискурс – речь, связный текст в совокупности с прагматическими, социокультурными и другими факторами; речь как целенаправленное социальное действие.

Инсталляция – установка программного продукта на ЭВМ.

Интеллект – мыслительная способность человека, уровень умственного развития, способность к рациональному познанию.

Интернет – всемирная информационная компьютерная сеть, связывающая между собой как пользователей компьютерных сетей, так и пользователей индивидуальных компьютеров, позволяющая обмениваться разнообразной информацией.

Интерфейс – система унифицированных связей и сигналов, посредством которых устройства вычислительной системы взаимодействуют друг с другом.

Информатика – 1) отрасль науки, изучающая общие свойства информации, а также вопросы, связанные с её накоплением, преобразованием, поиском, хранением и передачей с помощью компьютеров и других вычислительных средств; 2) сфера практического применения вычислительной техники».

Информация – 1) сообщение о фактах, событиях, о состоянии чего-либо; 2) совокупность сведений как объект хранения, переработки и передачи.

Искусственный интеллект – в лингвистике направление, которое объединяет исследования в области компьютерной лингвистики и машинного обучения.

Квантитативная лингвистика – наука, исследующая язык с помощью точных математических методов и компьютерных программ.

Когнитология – наука о знаниях, приемах и методах получения, обработки, хранения и использования знаний.

Коммуникация – 1) путь сообщения; 2) форма связи; 3) акт сообщения, связь между индивидами.

Компьютер – вычислительная машина, основные элементы которой (запоминающие, логические и др.) выполнены на электронных приборах.

Корпусная лингвистика – специальный раздел языкознания, который занимается проблемами разработки, создания и использования текстовых корпусов.

Корпусный менеджер – программа, предоставляющая удобный доступ к базе данных корпуса.

Криптография – наука о математических методах и алгоритмах преобразования данных для обеспечения их конфиденциальности, целостности и аутентификации.

Лексикостатика – один из методов сравнительной лингвистики, применяемый для сравнения процентного соотношения родственных слов разных языков и определения взаимосвязи между словами и языками.

Лингвистика – наука о естественном человеческом языке вообще и о всех языках мира как индивидуальных его представителях.

Лингвистический корпус – совокупность текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту и обеспеченных специализированной поисковой системой.

Логика – наука об общезначимых формах и средствах мысли (в том числе закономерностях мышления, методах рассуждений), необходимых для рационального познания в любой области знания, непротиворечивых выводов, формулирования понятий, построения умозаключений, гипотез, версий, теорий.

Макропроцессор – программа, техническое устройство, осуществляющее управление ЭВМ при помощи системы команд в соответствии с рабочим машинным языком.

Маршрутизатор – компьютер, объединяющий несколько сетей и предназначенный для определения маршрута передачи данных в те сети, для которых они предназначены.

Макрос – 1) средство замены одной системы символов другой; 2) последовательность команд, запускаемая одним нажатием клавиш на клавиатуре или кнопки на экране.

Мейнфрейм – универсальный компьютер высокого уровня, предназначенный для решения задач, связанных с интенсивными вычислениями и обработкой больших объемов информации.

Метакоммуникация – вторичная коммуникация, включающая косвенные сигналы о том, как предполагается интерпретировать фрагмент информации, коммуникация «по поводу коммуникации».

Моделирование – исследование каких-либо объектов на их моделях; построение моделей реально существующих предметов, явлений или процессов (живых организмов, инженерных конструкций, образцов одежды и т.п.)

Моделирование общения – направление компьютерной лингвистики, задачей которого является представление в формальном виде общение человека с персональным компьютером.

Модератор – человек, имеющий более широкие права по сравнению с обычным пользователем на общественных сетевых ресурсах.

Нарратив – самостоятельно созданное повествование о некотором множестве взаимосвязанных событий, представленное читателю или слушателю в виде последовательности слов или образов.

Носитель информации – материальный объект, предназначенный для хранения данных.

Пен-компьютер – портативный компьютер, в котором основным устройством ввода информации является перо (а не клавиатура и мышь).

Поиск – определенная последовательность действий, операций, целью которых является сбор и обработка необходимой информации.

Поисковая машина – комплекс программ, предназначенный для поиска информации.

Поисковая система – автоматизированный программно-аппаратный комплекс с веб-интерфейсом, предоставляющий возможность поиска информации в Интернете.

Порождающая грамматика – формализм генеративной лингвистики, связанный с изучением синтаксиса.

Прикладная лингвистика – направление в языкознании, занимающееся разработкой методов решения практических задач, связанных с использованием языка.

Релятор – символ или слово, которые используются для различения значений многозначного термина.

Символический искусственный интеллект – собирательное название для всех методов искусственного интеллекта, основанных на высокоуровневом «символическом» (человекочитаемом) представлении задач, логики и поиска.

Синсет – один из ключевых элементов в лексической базе данных WordNet и подобных ей системах.

Структурная лингвистика – научное направление, которое рассматривает язык как знаковую систему с четко выраженными структурными элементами.

Структурное литературоведение – направление в науке, задачей которого является обнаружение, описание и объяснение структуры мышления, лежащей в основе культуры прошлого и настоящего.

Тег – ассоциированное ключевое слово, относящееся к какой-либо информации (текст, фото, видео, закладки браузера и другие файлы).

Тезаурус – тип словаря, в котором все слова данного языка представлены максимально полно и исчерпывающим перечнем их употребления в тексте.

Телетайп – приемопередающий буквопечатающий телеграфный аппарат, который применяется для передачи сообщений по каналам связи, а также в качестве терминала в устройствах вычислительной техники

Теоретическая лингвистика – наука, изучающая объективное состояние языка, его историю, сложившиеся в нем закономерности и т.д.

Терминал – устройство для ввода и вывода информации (клавиатура, дисплей, принтер и др.)

Формализм – операция замены смысловых содержательных описаний некоторых явлений, процессов, состояний математическими структурами.

Этимон – первоначальное значение и форма слова, от которого произошло слово современного языка.

Applied linguistics – лингвистическое направление, задачей которого является разработка способов и приемов обучения иностранному языку, включая методику преподавания иностранного языка, особенности описания грамматики в учебных целях и т. п.

Hardware – область информатики, занимающаяся вопросами аппаратного обеспечения.

Lingware – область информатики, занимающаяся вопросами лингвистического обеспечения.

Software – область информатики, занимающаяся вопросами программного обеспечения.

ЛИТЕРАТУРА

Основная

1. Направления и методы лингвистических исследований: учебное пособие для студентов учреждений высшего образования второй ступени (магистратура) по лингвистическим специальностям / Е.А. Красина [и др.]; под ред. Е.А. Красиной, В.А. Масловой. – Минск: РИВШ, 2020. – 189 с.

2. Рогалев, А.Ф. Общее языкознание. Концепции языка: практическое пособие для студентов / А.Ф. Рогалев; Учреждение образования «Гомельский государственный университет имени Франциска Скорины». – Гомель: ГГУ им. Ф. Скорины, 2021. – 45 с.

3. Фефилов, А.И. Языкознание. Общая теория и история: учебник / А.И. Фефилов; М-во науки и высшего образования Российской Федерации, Федеральное государственное бюджетное образовательное учреждение высшего образования «Ульяновский государственный университет». – Москва: ФЛИНТА, 2022. – 255 с.

Дополнительная

1. Марчук, Ю.Н, Компьютерная лингвистика: учебное пособие / Ю.Н. Марчук. – М.: АСТ: Восток – Запад, 2007. – 317 с.

2. Голуб, О.Ю. Теория коммуникации: учеб. для студентов вузов, обучающихся по напр. «Реклама и связи с общественностью» / О.Ю. Голуб. – М.: Дашков и К, 2014. – 388 с.

3. Щипицы, Л.Ю. Информационные технологии в лингвистике: учеб. пособие / Л.Ю. Щипицына. – М.: Флинта: Наука, 2013. – 128 с.

Учебное издание

ИНФОРМАЦИОННАЯ ЛИНГВИСТИКА ДЛЯ СПЕЦИАЛЬНОСТИ 7-06-0232-01 ЯЗЫКОЗНАНИЕ

Учебно-методический комплекс по учебной дисциплине

Составитель

ГЕНКИН Владимир Максимович

Технический редактор *Г.В. Разбоева*

Компьютерный дизайн *Л.В. Рудницкая*

Подписано в печать 16.12.2024. Формат 60x84 $^{1}/_{16}$. Бумага офсетная.
Усл. печ. л. 4,13. Уч.-изд. л. 3,80. Тираж 9 экз. Заказ 187.

Издатель и полиграфическое исполнение – учреждение образования
«Витебский государственный университет имени П.М. Машерова».

Свидетельство о государственной регистрации в качестве издателя,
изготовителя, распространителя печатных изданий
№ 1/255 от 31.03.2014.

Отпечатано на ризографе учреждения образования
«Витебский государственный университет имени П.М. Машерова».
210038, г. Витебск, Московский проспект, 33.