МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования
«Тихоокеанский государственный университет»

Социально-гуманитарный факультет

Кафедра русской филологии

Направление 031100. 68 – Лингвистика (магистратура)

2 курс

Учебно-методические материалы

по дисциплине

Компьютерные технологии в лингвистических исследованиях

Автор-разработчик:

д. филол.н, проф. Крапивник Л.Ф.

Хабаровск

Тексты лекций

Тема 1. Прикладная лингвистика как научное направление

Прикладная лингвистика – это научное направление в языкознании, которое ориентировано на лингвистическое обеспечение информационных систем разных типов, т.е. на прикладные задачи – машинный перевод, компьютерное обучение иностранным языкам и т.п. От теоретической лингвистики она отличается тем, что:

- изучает не язык в его состоянии (т.е. системе), а язык в действии (т.е. в общении);

- решает конкретную прикладную задачу, создавая языковые модели, и при этом не претендует на объяснение фактов языка (как теоретическая лингвистика);

- ориентирована на конкретные подъязыки (т.е. на выборочные знания о языке), а не на весь язык в целом.

Прикладная лингвистика использует автоматическую обработку языка в его устной и письменной формах, т.е. она связана с широким использованием ЭВМ в процессе лингвистического анализа.

Большое внимание прикладная лингвистика уделяет систематизации лингвистического материала и их классификации. Поэтому развитие прикладной лингвистики и ее достижения позволили создать большие банки хранения лингвистической информации (картотеки и словари), которыми пользуются специалисты по гуманитарным наукам

В связи с этим основной особенностью прикладной лингвистики является использование новых методов анализа языка и новых приемов его описания. В частности, в прикладной лингвистике широко используются методы математики, например, статистический метод и метод

моделирования, которые помогают автоматизировать процесс лингвистического исследования.

Ядром прикладной лингвистики является структурная и математическая лингвистика. Их задачей является разработка и совершенствование структурных и формальных методов анализа и описания языка.

Тема 2. Компьютерная лингвистика как одно из направлений прикладной лингвистики

Одним из направлений в прикладной лингвистике является компьютерная лингвистика. Ее цель – разработка методов, технологий и конкретных систем, обеспечивающих общение человека с ЭВМ на естественном или ограниченном естественном языке.

При моделировании функционирования языка в тех или иных условиях, ситуациях и сферах компьютерная лингвистика ориентируется на использование компьютерных инструментов – программ, компьютерных технологий организации и обработки данных. Таким образом, компьютерная лингвистика как прикладная дисциплина выделяется прежде всего по инструменту – т.е. по использованию компьютерных средств обработки языковых данных.

Важнейшие направления компьютерной лингвистики следующие:

- создание систем обработки естественного языка (например, систем обработки связного текста);

- разработка информационно-поисковых систем (документальных, т.е. в которых хранятся тексты, и фактографических, т.е. в которых хранятся факты, представленные не только в текстовой форме, то и в форме таблиц, формул и т.п.);

- создание гипертекстовых систем (т.е. множества текстов со связывающими их отношениями);

- разработка компьютерных технологий составления и эксплуатации словарей.

В рамках компьютерной лингвистики создаются специальные программы – базы данных, компьютерные картотеки, программы обработки текстов, которые позволяют в автоматическом режиме формировать словарные статьи, хранить словарную информацию и обрабатывать ее. Компьютерная лингвистика занимается также и машинным переводом.

Тема 3. Компьютерные словари

Компьютерные словари стали сегодня непременной частью личной библиотеки любого интеллигентного человека, в том числе и ученого-лингвиста. **Словари** и сама концепция электронной книги оказались как будто созданными друг для друга. Поэтому на сегодняшний день в магазинах имеется неплохой ассортимент компьютерных словарей иностранных языков.

Примерно за десять последних лет компьютерный словарь научился:

- сам находить нужное слово,

- заговорил,

- уместился в удобном компакт-диске,

- начал активно помогать пользователю учиться.

Вторую жизнь в электронном виде получили многие известные словари.

Новый Большой англо-русский словарь (НБАРС) объемом 250.000 слов под редакцией академика Ю.Д. Апресяна был переведен на компьютер компанией МультиЛекс в 1996 году и с тех пор неоднократно совершенствовался. Кроме этого, в программной оболочке МультиЛекс имеются: англо-русский и русско-английский словарь под редакцией О.С. Ахмановой и Е.А.М. Уилсон (40.000 слов), англо-русский словарь В.К. Мюллера (60.000 слов), русско-английский словарь под редакцией А.И. Смирницкого (55.000 слов) и коллекции специальных словарей.

Многие из популярных компьютерных словарей интересны тем, что в них можно найти географические справки, очерки о явлениях культуры и даже имена и биографии известных людей.

Компьютерный словарь может выполнять множество **служебных функций**.

1). **Автоматически отыскивать по запросу словарную статью.** Это произошло примерно в 1995-1996 годах. Сегодня работа в любом из электронных словарей начинается с "окна поиска" - строки, где достаточно набрать слово, которое вы ищете. Именно эта не очень сложная функция экономит время при использовании электронных словарей.

2). **Запоминать страницы, которые вы открывали, и возвращаться** по команде "Назад" туда, где вы побывали только что; следующим шагом можно вернуться туда, где были еще раньше, и так идти по своим следам, в принципе, сколь угодно долго.

3). Практически все словари позволяют **"выписывать" нужные слова** в "блокноты" или "**ставить закладки**".

3). **Выполнять функцию "гипертекст".** Например, словарь Апресяна в книжном варианте - англо-русский. Но электронная версия имеет дополнительную возможность: "отметив" мышкой на экране любое русское

слово (или набрав его в окне поиска), мы получаем полную подборку словарных статей, где слово встречается. При переводе с русского на английский такая функция даже полезнее обычного русско-английского словаря, потому что она позволяет полнее видеть контекст.

В толковом словаре *Collins* смысл гипертекста иной - не случайно на его обложке стоит лозунг "Думай и говори по-английски". Толковый словарь - спутник ученика, уже перешедшего к полному погружению в изучаемый язык. Если в толковании нового слова встретилось еще одно или несколько непонятных слов, достаточно щелкнуть на любом из них мышкой, и вы переходите уже к его словарной статье. Легко и вернуться к первому слову. Так технология этого словаря позволяет прощупать и прочувствовать смысл нового слова.

У *Longman*'а гипертекст работает так же, только в урезанном объеме. Щелкнуть мышкой имеет смысл только на том слове, что выделено цветом - а таких обычно всего два-три в словарной статье.

Аналогичный эффект дает функция англо-русско-английского "обратного перевода" в *Partner*. Там после выбора русского слова на экране появляется набор его возможных английских синонимов без комментариев. Берем один - видим спектр его значений уже на русском. Вновь выбираем один перевод и так, в принципе, до бесконечности. "Обратный перевод" наглядно демонстрирует, что полных синонимов (кроме специальных терминов, конечно) в разных языках почти нет. Именно в этом и состоит "изюминка" процесса изучения новых слов…

4). **Выполнять функцию текстового редактора**. Так, окно поиска словарей *Collins* позволяет помещать для пословного разбора целые фрагменты текста, с которым вы работаете. В аналогичное окно *Lingvo* при определенной сноровке можно перетаскивать мышкой слова непосредственно из окна *Word* или другого приложения *Windows*.

5). Некоторые словари предлагают и другие возможности. *Partner* пытается **найти в словарной базе даже слова, написания которых вы не знаете**! Для этого в окне поиска надо набрать слово так, как вы его услышали - *DOTA* вместо *daughter*, *PIS* вместо *peace* и так далее - а программа, принимая в расчет возможные орфографические ошибки, постарается подобрать правильные варианты. После этого можно прослушать предложенные слова, выбрать похожее на то, что вы ищете, и посмотреть перевод.

6). В 1997-98 годах ведущие разработчики начали **озвучивать** свои словари, а примерно с 2000 года компьютерный словарь обязан быть **говорящим**. Где-нибудь на его экране обычно размещена кнопочка с изображением репродуктора; щелкнув мышкой по ней, **можно услышать, как звучит выбранное слово.**

7). Ряд словарей снабжен системой **быстрого заучивания новых слов.** Само собой, такая система еще не превращает словарь в учебник. Однако любой курс английского построен так, что к каждому уроку дается пригоршня новых слов - и выучить их с интерактивным словарем намного проще.

8). В компьютерном словаре пользователь **может формировать "блокноты**". Чтобы внести слово в "блокнот", достаточно щелкнуть по нему правой кнопкой мыши. Блокноты открываются, копируются, редактируются и удаляются как обычные компьютерные файлы.

9). Многие словари дают возможность **практиковаться в произношении выбранных слов**, позволяют ученику не только записывать и прослушивать собственное произношение, но и сравнить график (осциллограмму) собственной речи с дикторской.

Тема 4. Системы компьютерного перевода

Этапы развития компьютерного перевода

Первые программы машинного перевода появились в 50-х годах, через несколько после рождения компьютера. В это время машинный перевод был объектом научных исследований, т.е. изучались возможности машинного перевода текстов. В это время компьютерный перевод не получил еще широкого распространения. Этому две причины:

- дороговизна времени работы компьютера,

- невозможность его оперативно использовать, т.к. в это время было коллективное пользование ресурсами компьютера.

В начале 80-х годов компьютеры начали завоевывать мир, т.е. получили широкое распространение. Время их работы подешевело и доступ к ним можно было получить в любую минуту. А значит, машинный перевод стал **экономически выгодным.**

В эти и последующие годы стали более совершенными программы компьютерного перевода. Это позволило достаточно точно переводить многие виды текстов и активно использовать программы компьютерного перевода.

Однако некоторые проблемы машинного перевода остались до сих пор нерешенными. Поэтому современный компьютер не может дать полноценный перевод. Например, он не всегда может понять содержание текста в полной мере. Кроме того, он не понимает языковых нюансов, намеков в тексте, того, что называется тонкой игрой слов.

Принципы работы компьютерного переводчика.

Компьютерный переводчик работает следующим образом:

- предложение расчленяется на части речи,

- в нем выделяются стандартные конструкции,

- слова и словосочетания переводятся по находящимся в памяти машины словарям,

- затем переведенные части речи собираются по правилам другого языка.

Трудности машинного перевода.

Трудности машинного перевода связаны с особенностями функционирования языка. Они могут быть разного характера – стилистические, лексические, синтаксические, страноведческие, художественные.

1). Он не всегда учитывает значения, которые может иметь слово в разных стилях речи.

2) Делает ошибки в переводе слов в **устойчивых словосочетаниях** и **фразеологизмах,**

3) Не учитывает «красоты языка», т.е. дополнительные смыслы, которые возникают при изменении порядка слов.

4) Не может определить, как изменяется значение слова в зависимости от контекста.

Особенности автоматического перевода технического текста и литературного текста.

Перевод технического текста отличается от перевода литературного текста. При техническом переводе важно знать принятые за рубежом стандарты обозначений тех или иных понятий. При литературном переводе требуется получить текст, по художественной ценности максимально близкий к оригиналу.

При переводе технических текстов, если правильно выбрать словарь по специальности, к которой относится текст, то получается вполне удовлетворительный результат. Этот перевод почти не требует помощи человека. Если компьютер используется для перевода литературных текстов, то получается черновой вариант текста, так называемый **подстрочник.** Подстрочник превращается в произведение искусства человеком, который слабо знает язык оригинала, но является хорошим литературным редактором. При переводе художественных текстов компьютер пока не может заменить переводчика.

Современные переводные программы

Современные компьютерные переводные программы постоянно совершенствуются.

1). Современные системы машинного перевода обязательно имеют средства редактирования текстов.

2). Создаются системы компьютерного перевода с элементами искусственного интеллекта (в них имитируется мыслительная деятельность человека).

В мире существует очень много программ машинного перевода. В России наиболее распространены системы Stylus (фирма «ПроМТ») и ПАРС (фирма «Лингвистика 93»). **Stylus** предназначена для профессионального перевода больших объемов информации (это очень дорогая программа).

Имеется **Система ПАРС.** Она достаточно удобна для бытового использования и доступна (т.к. стоит недорого). К ней имеется большой набор словарей по различным темам: вычислительная техника, медицина, химия и т. Д). Эта система хорошо работает в среде Windows 3.1 и более поздних версий. Эта система имеет некоторые особенности работы.

А) Если перевод осуществляется впервые после запуска программы, перед его началом потребуется указать используемые словари. Словари выбираются в зависимости от стиля и тематики текста.

Б) системы машинного перевода могут ошибаться из-за наличия в тексте **сокращений, заканчивающихся точкой**. Сокращения будут перенесены в текст без перевода, и их нужно перевести вручную.

В) в переводимом тексте **должны отсутствовать переносы.**

Сейчас наблюдается повышение интереса к системам машинного перевода в связи с развитием Internet. Доминирует там английский язык. Для облегчения просмотра страниц Internet на незнакомом пользователю языке появились дополнительные системы, которые немедленно переводят нужные фрагменты Webстраницы.

<u>Тема 5.</u> Математическое моделирование в лингвистике: метод статистического анализа

Статистические данные – это количественные сведения о какой-либо совокупности объектов, которые имеют общие признаки, способные изменяться качественно и количественно.

Статистический метод – это комплекс приемов и принципов, согласно которым производятся сбор, систематизация, обработка и интерпретация статистических данных с целью получения научных и практических выводов.

Математическое содержание приемов и принципов статистического метода образует **математическая статистика**, которая является отраслью прикладной математики. Основными категориями математической

статистики являются *вероятность, частота, случайная величина, выборка, корреляция* и др.

В традиционной статистике различаются две группы методов:

- описательные методы,

- методы оценивания.

Задача описательных методов – представить исходные данные в компактной и наглядной форме (в виде таблиц, графиков) и описать эти данные с помощью разного рода статистик (мер связи, мер концентрации, мер центральной тенденции).

Методы оценивания распадаются на две группы: методы оценивания неизвестных параметров распределения и методы проверки статистических гипотез.

Статистический метод – это универсальный метод познания действительности. Он имеет несомненные преимущества по сравнению с другими научными методами, например, такие как объективность и беспристрастность, строгость и процедурность. Поэтому метод статистического анализа в разных науках активно используется для компактного представления, анализа, обобщения и интерпретации данных наблюдения и эксперимента.

Использование статистических методов в лингвистике не является простой процедурой. Эти две науки требуют «приспособления» друг к другу. Адаптация статистического метода к решению филологических проблем осуществляется в двух направлениях:

- лингвистическое переосмысление статистических категорий (*выборка, корреляция* и др.);

- статистическое переосмысление лингвистических категорий («язык», «речь», «текст» и др.).

Использование метода статистического анализа требует от лингвиста владения как лингвистической проблематикой, так и аппаратом математической статистики.

Лекция 6. Лингвостатистический метод и его особенности

В основе создания лингвостатистического метода лежит представление о том, что наука достигает совершенства лишь тогда, когда использует точные математические методы.

Теоретическое обоснование методов количественного анализа и создание алгоритмов их практического применения в лингвистике – это предмет особой отрасли науки о языке, получившей название **лингвостатистики.**

Суть лингвостатистического метода заключается в установлении количественных изменений, вызывающих качественные преобразования языковых явлений. Благодаря использованию математических методов исследования языка в рамках лингвостилистики было выявлено, что частота появления тех или иных языковых элементов в речи подчиняется определенным статистическим законам (закономерностям). Это позволяет на основе статистических данных сформулировать определенные закономерности функционирования единиц языка и построения текста.

Лингвостатистический метод широко применяется в современной лексикологии и стилистике. Он используется для изучения как явлений языка, так и явлений речи. Например, с помощью лингвостатистического метода лингвисты изучают количественные характеристики словарного состава в разных стилевых и авторских разновидностях речи. В

результате лингвостатистического изучения языка появились **частотные словари**.

Количественное описание подъязыков науки и техники используется для автоматической обработки языковой информации (создания информационно-поисковых систем), а также в методике преподавания языков.

Вопросы
практических занятий

Практическое занятие 1.

Прикладная лингвистика как научное направление

(2 часа)

Вопросы для практического занятия.

1. Какое научное направление в языкознании называется прикладной лингвистикой?

2. На решение каких прикладных задач ориентирована прикладная лингвистика?

3. Как отличается прикладная лингвистика от теоретической лингвистики по предмету исследования?

4. Как отличается прикладная лингвистика от теоретической лингвистики по задачам исследования?

5. Как отличается прикладная лингвистика от теоретической лингвистики по материалу исследования?

6. Как отличается прикладная лингвистика от теоретической лингвистики по способам исследования языкового материала?

7. Как отличается прикладная лингвистика от теоретической лингвистики по методам исследования языкового материала?

8. Как отличается прикладная лингвистика от теоретической лингвистики по результатам исследования языкового материала?

9. Какие разделы выделяются в прикладной лингвистике?

10. Дайте сопоставительную характеристику прикладной и теоретической лингвистики, заполнив таблицу.

Сравнительная характеристика прикладной и теоретической лингвистики

Лингвистическое направление	**Предмет изучения**	**Задачи**	**Материал изучения**	**Способы исследования**	**Методы исследования**	**Результаты исследования**	**Разделы**
Теоретическая лингвистика							
Прикладная лингвистика							

Практическое занятие 2.

Компьютерная лингвистика как одно из направлений прикладной лингвистики

(2 часа)

Вопросы для практического занятия.

1. Какое научное направление в языкознании называется компьютерной лингвистикой?

2. Какой научный инструментарий использует компьютерная лингвистика?

3. Назовите важнейшие направления исследований компьютерной лингвистики.

4. Какие научные программы создаются рамках компьютерной лингвистики?

5. Какие научные результаты получены в рамках компьютерной лингвистики?

6. Охарактеризуйте особенности компьютерной лингвистики как одного из направлений прикладной лингвистики, заполнив таблицу.

Компьютерная лингвистика как одно из направлений прикладной лингвистики

Цели исследования	**Научный инструмен-тарий**	**Направления исследований**	**Научные программы**	**Научные результаты**	**Отличительные черты**

Практическое занятие 3.

Компьютерные словари

(3 часа)

Вопросы для практического занятия.

1. Почему компьютерные словари стали сегодня непременной частью личной библиотеки любого интеллигентного человека, в том числе и ученого-лингвиста?

2. Назовите отличия компьютерных словарей от обычных словарей.

3. Какие известные словари получили вторую жизнь в электронном виде?

4. Объясните следующие понятия и термины, имеющие отношения к компьютерным словарям:

- окно поиска,

- словарная статья,

- блокнот,

- гипертекст,

- текстовый редактор.

5. Опишите особенности действий, которые применяются при использовании компьютерного словаря:

- поиск словарной статьи,

- поставить закладку,

- обратный перевод.

6. Перечислите и охарактеризуйте служебные функции компьютерных словарей.

7. Почему компьютерные словари могут быть использованы при изучении иностранных языков.

8. Какие служебные функции компьютерных словарей используются при изучении иностранных языков?

9. Охарактеризуйте возможности компьютерного словаря, которым Вы пользуетесь.

Практическое занятие 4.

Системы компьютерного перевода

(4 часа)

Вопросы для практического занятия.

1. Расскажите об этапах развития компьютерного перевода.
2. Охарактеризуйте современный этап развития компьютерного перевода.
3. Расскажите о принципах работы компьютерного переводчика.
4. Назовите трудности компьютерного перевода и охарактеризуйте их на конкретных примерах.
5. Расскажите об особенностях перевода технического текста.
6. Расскажите об особенностях перевода литературного текста.
7. Расскажите о современных компьютерных переводных программах.
8. Расскажите об особенностях работы современных компьютерных переводных программ.

Практическое занятие 5.

Математическое моделирование в лингвистике: метод статистического анализа

(2 часа)

Вопросы для практического занятия.

1. Раскройте содержание следующих понятий и терминов:

- статистические данные,
- статистический метод,

- математическая статистика.

2. Раскройте содержание следующих основных категорий математической статистики:

- вероятность,
- частота,
- случайная величина,
- выборка,
- корреляция.

3. Какие группы методов выделяются в традиционной статистике?

4. Охарактеризуйте описательные методы.

5. Охарактеризуйте методы оценивания.

6. В чем преимущества статистического метода по сравнению с другими научными методами?

7. Для чего используется метод статистического анализа в разных науках?

8. Каковы главные особенности использования статистических методов в лингвистике?

9. В каких направлениях осуществляется адаптация статистического метода к решению филологических проблем?

Практическое занятие 6.

Лингвостатистический метод и его особенности

(2 часа)

Вопросы для практического занятия.

1. Что лежит в основе создания лингвостатистического метода?

2. Что является основным направлением деятельности особой отрасли науки о языке, получившей название лингвостатистики?

3. В чем заключается суть лингвостатистического метода?

4. Какие основные научные результаты были получены благодаря использованию статистических методов исследования языка?

5. В каких сферах лингвистического знания применяется лингвостатистический метод?

6. Для чего используется количественное описание подъязыков науки и техники?