**Обзор современных направлений в прикладной лингвистике**

**Введение**

**Прикладная лингвистика** -наука о функционировании языка и моделировании языковых функций с точки зрения приложения лингвистической теории в разных нелингвистических областях знаний.

Прикладная лингвистика - общий термин, использующийся для обозначения деятельности “по приложению научных знаний об устройстве и функционировании языка в нелингвистических научных дисциплинах и в различных сферах практической деятельности человека, а также теоретическое осмысление такой деятельности” (Баранов: 7). В западной традиции термин прикладная лингвистика “applied linguistics, angewante Linguistik” связывается с разработкой методик преподавания языков и теорию перевода. В российской лингвистической традиции этот термин трактуется гораздо шире. Кроме направлений, связанных с теорией и практикой преподавания языков и переводов, к нему относят, во-первых, все виды деятельности, так или иначе связанные с автоматической обработкой естественного языка (Natural language processing), а во-вторых, деятельности, целью которых является регламентация и нормализация языка: лексикография, лингвистическая экспертиза, терминоведение и др.

Прикладная лингвистика как самостоятельная научная дисциплина возникла сравнительно недавно (приблизительно к 1920-м годам). Однако задачи, связанные с приложением лингвистических знаний в различных сферах человеческой деятельности, включая моделирование процесса познания, обучение языку, создание нормативных языковых описаний и т.п. стояли перед языкознанием уже давно. Под "шапкой" данной области исследования очень часто объединяют множество, на первый взгляд, разнородных направлений и приложений. С одной стороны, отдельные области прикладной лингвистики рассматриваются как части дисциплин в рамках сформировавшегося недавно научного направления, называемого общим термином Computer Science. С другой стороны, лингвистические модели исследуются и применяются в сугубо гуманитарных областях таких, как практика преподавания языка, теории речевого воздействия, политической лингвистике. Согласно А.Н.Баранову все эти направления можно объединить в том смысле, что в них "изучаются и разрабатывают способы оптимизации функционирования языка".

Не существует единой точки зрения на состав основных направлений прикладной лингвистики. Обычно выделяют следующие направления, связанные с изучением языка:

1. **лексикография** — теория и практика составления словарей;
2. **лингводидактика** — наука о разработках методик обучения иностранному языку;
3. **терминоведение** — наука об упорядочении и стандартизации научно-технической терминологии;
4. **переводоведение** — теория перевода
5. **квантитативная лингвистика** (изучение частотных свойств языковых единиц и их связь с другими свойствами, такими как фонетическая сложность, морфологическая сложность, многозначность, возраст).

Основные направления прикладной лингвистики, связанные с практическими приложениями:

1. **Компьютерная лингвистика** (computational linguistics)
2. **Лингвистическая экспертиза** (например, в судебной практике)
3. **Политическая лингвистика** (анализ политического дискурса)

Одной из важных областей применения лингвистических знаний в других областях является **прикладная фонетика**. Она занимается практическим применением фонетических знаний, накопленных в лингвистике. Ее разделы отражают многообразие использования языка в человеческой жизни: к прикладной фонетике относятся такие разные дисциплины, как *методика преподавания* фонетики конкретного языка, *фонетическая риторика* – приемы выразительной речи, *орфоэпия*, устанавливающая различные произносительные нормы. Важное применение фонетика находит также в *логопедии*, *речевой дефектологии* и лечении болезней, вызванных повреждением речевых функций.

В последнее время прослеживается тенденция к использованию термина “прикладная лингвистика именно в западном значении”. Так, Национальное общество прикладной лингвистики занимается именно проблемами лингводидактики ([www.nopril.ru](http://www.nopril.ru/)). Технологии анализа естественного языка, моделирования когнитивных процессов понимания и языкового взаимодействия и извлечения информации из текстов объединяются общим термином “Компьютерная лингвистика” (вычислительная лингвистика, computational linguistics).

Несмотря на то, что методы, применяемые в различных направлениях прикладной лингвистике разнообразны, можно выделить общие характерные признаки:

1. ведущая роль метода моделирования;
2. экспериментальный характер прикладных методик
3. комплексное сочетание разных наук.

В России термин «прикладная лингвистика» получил широкое распространение в 1950-х годах. Это обстоятельство связано с появлением первых компьютерных систем автоматической обработки текстовой информации (машинного перевода, автоматического реферирования и др.). В русскоязычной литературе распространен подход, при котором термин «прикладная лингвистика» имеет то же значение, что и «компьютерная лингвистика», «вычислительная лингвистика», «автоматическая лингвистика», «инженерная лингвистика.

**Компьютерная лингвистика как одно из актуальных направлений современной прикладной лингвистики**

В рамках направления компьютерная лингвистика в широком смысле слова объединяются как компьютерное моделирование с целью проверки адекватности теоретических формально-математических моделей языка и его отдельных уровней, так и применение различных математических методов для решения конкретных прикладных задач в различных системах обработки информации.  
Среди современных направлений компьютерной лингвистики можно выделить следующие (см. презентацию **комп\_линг**):

1. [Машинный перевод](http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127458:article)
2. [Речевые технологии](http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127472:article) (в частности, автоматическое распознавание речи, ASR)
3. [Лингвистическое обеспечение информационного посика](http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127457:article)
4. Автоматическое извлечение данных (Data Mining)
5. Автоматическое реферирование текстов
6. Создание **электронных лексикографических ресурсов** (словарей, онтологий)
7. [Корпусная лингвистика](http://www.lomonosov-fund.ru/enc/ru/encyclopedia:01210:article) (создание и использование электронных корпусов текстов)
8. Разработка **вопросно-ответных систем**

Часть данных направлений, а также разработка систем автоматического анализа на разных языковых уровнях: морфологических анализаторов (парсеров), систем автоматического синтаксического анализа и т.п. часто выделяют в отдельную область - **автоматическая обработка естественного языка** (*Natural Language Processing, NLP* — общее направление искусственного интеллекта и компьютерной лингвистики.

История и основные тенденции развития компьютерной лингвистики

Компьютерная лингвистика возникает как одно из направлений искусственного интеллекта в 50-х годах 20-го века и основные этапы его развития во многом соотносятся с эволюцией подходов и методов в искусственном интеллекте в целом.

Толчком к развитию искусственного интеллекта и компьютерной лингвистики в частности послужило создание первых ЭВМ в 40-х годах, и, в частности, успешное использование их во время второй мировой войны. Первый этап фундаментального осмысления задач, стоящих перед разработчиками искусственного интеллекта, связывают с классической статьей **Тьюринга, (Alan Turing 1950)**, которая начинается с рассмотрения философски значимого вопроса “может ли машина мыслить” ('Can machines think?'). Тьюринг предположил, что о создании искусственного интеллекта можно будет говорить в том случае, если воображаемый компьютер сможет успешно пройти следующий тест: человек, находящийся в другой комнате и задающий вопросы не может отличить полученные ответы другого человека от ответов машины. Этот тест получил название теста Тьюринга и был подвергнут широкому обсуждению. Важно отметить, что в основе теста лежат проблемы, относящиеся к области автоматической обработки естественного языка и моделирования языкового взаимодействия.

С возникновением нового поколения компьютеров, а также первых языков программирования, начались эксперименты с [машинным переводом](http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127458:article). В 1960-х годах подобные исследования начались и в СССР. Основной областью разработок стали системы [машинного перевода](http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127458:article), наиболее известная из которых –ЭТАП – разрабатывается и по сей день в «Лаборатории компьютерной лингвистики» Института Проблем Передачи Информации.

Первые разработки в области компьютерной лингвистики были связаны с попытками моделирования языковых способностей человека. Проводились эксперименты по созданию систем машинного перевода и интеллектуальных систем, моделирующих языковое поведение человека (разработка систем общения человека с компьютером, систем понимания естественного языка). Первая публичная демонстрация машинного перевода (так называемый Джорджтаунский эксперимент) состоялась в 1954 году. Однако достаточно скоро стало ясно, что в решении таких задач ни на одном из данных направлений не удается достичь быстрых успехов.

Главный фокус в применении компьютера для анализа естественного языка сместился, с одной стороны, на создание экспериментальных систем, моделирующих коммуникативное человеко-машинное взаимодействие в ограниченной предметной области, с другой стороны, на создание реальных приложений, в той или иной степени использующих ограниченный естественный язык, например, создание Информационно-поисковых систем. В конце 80-х гг. связи с развитием сети Интернет и Всемирной паутины объем доступных в электронном виде текстов возрос в разы. Это привело к качественному скачку в технологиях информационного поиска. Возникли совершенно новые задачи и технологии обработки текстов на естественном языке. Необходимость быстрой обработки огромного потока неструктурированных данных привела к тому, что основной интерес в сфере автоматической обработки естественного языка переместился на разработку статистических методов. Методами статистического анализа решались такие задачи, как выделение тематически значимых элементов из текста, автоматическое разбиение текстов на кластеры, объединенные одной тематикой, например, автоматическое объединение новостных сообщений, относящихся к одному и тому же событию. Такие задачи как распознавание и генерация речи, создание поисковых систем до настоящего времени решаются с минимальным участием лингвистов. Это обусловлено использованием при решении вышеупомянутых задач в основном статистических методов, методов машинного обучения. Однако новый количественный скачок в объеме доступного в сети контента (любое информационно значимое наполнение информационного ресурса (например, веб-сайта)) привел к тому, что на первой план вышла проблема автоматического извлечения релевантной информации и представления ее пользователю в структурированном виде, упрощающем ее дальнейшую обработку. Для решения задач такого класса оказался востребован когнитивный анализ текста, невозможный без более глубокого лингвистического анализа (не только на уровне морфологии, но и на уровне синтаксиса и семантического преставления текстов) а также без моделей представления знаний и соотнесения знаний со способом их языкового выражения (разработки **лингвистических онтологий** и **лексикографических ресурсов**). “Интернациональный” характер сети привел к повышенному интересу к мультиязычным системам, задача создания систем машинного перевода вновь оказалась среди наиболее актуальных задач прикладной лингвистики. Современные поисковые машины в условиях жесткой конкуренции вынуждены предлагать пользователю все новые сервисы интеллектуального анализа текстов: автоматический анализ новостей: автоматическое объединение текстов, поступающих от новостных агентств в “новостные” сюжеты, определение “тональности” публикации, снабжение текста переводом на язык пользователя (ср. высплывающее окно перевода в Google) и др. Таким образом, возникшие за последние несколько лет огромные электронные текстовые ресурсы из самых различных областей знаний и практик поставили традиционную задачу лингвистики и филологии - анализ текста - на совершенно новый уровень: Для того, чтобы современный специалист мог эффективно справляться с такими большими информационными потоками, необходима их предварительная содержательная обработка. Оказалось, что большинство современных информационных технологий немыслимы без лингвистического компонента, обеспечивающего анализ текста на всех языковых уровнях:

1. фонетики (для систем обработки устной речи),
2. морфологии, синтаксиса (например, для современных поисковых систем)
3. семантики и лексики (например, для создания современных электронных словарных систем и семантического веба)

Лингвистические компоненты автоматического анализа текста

Практически любая современная система анализа текста, включая поисковые машины, осуществляющие поиск документов в сети Интернет, содержит те или иные модули автоматического лингвистического анализа.. Необходимыми этами лингвистического анализа практически в любой современной системе являются:

1. **токенизация** (Tokenization) (выделение слов и границ предложений);
2. **морфологический анализ** (приведение всех форм слова к одной словоформе) (см. презентацию **морфологический модуль**).

В зависимости от конкретных задач некоторые системы включают также и другие модули:

1. **модуль синтаксического анализа (синтаксический парсер)**, в задачи которого входит представление предложения в виде синтаксической структуры, например, дерева зависимостей или дерева непосредственных составляющих, см. **Синтаксис)** или ч**астичного синтаксического анализа (shallow parsing)**, либо модуль выделения в тексте отдельных словосочетаний **(chunker)**.(см. презентацию **синтаксический модуль**)
2. модуль семантического анализа, который, устанавливает семантические отношения между словами в тексте, объединяет различные языковые выражения, относящиеся к одному и тому же понятию и т.п. Работа семантического модуля обеспечивается с помощью различных лексикографических ресурсов, включая инфомарционно-поисковые тезаурусы и лингвистические онтологии (см. презентацию семантический модуль)
3. модуль [разрешения анафоры](http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127469:article) (anaphora resolution) и др.

На современном этапе технологии создания автоматических морфологических и синтаксических анализаторов (парсеров) достаточно хорошо разработаны. Принципиальным вопросом при разработке парсеров является выбор базовой алгоритмической модели, лежащей в основе. В синтаксическом анализе широко используются **порождающие грамматики**, в морфологическом - чаще представление в виде **конечных автоматов**. Что касается методов анализа, то на ранних этапах такие системы строились в основном на алгоритмической основе, сейчас при создании парсеров широко используются различные статистические методы, применяются методы распознавания образов (методы машинного обучения = методы самообучения), в частности, различные статистические модели, нейронные сети, так называемый bootstrapping. В основе перечисленных техноогий лежит следующий принцип: на вход системе подается некоторый “обучающий” корпус лингвистически размеченных данных (т.е. снабженных той или иной лингвистической информацией, ср. например, корпус, где каждому слову приписана его морфологическая характеристика). Анализируя такие обучающие данные, система сама извлекает оптимальные параметры для математической модели, лежащей в ее основе. Дальше эти параметры используются уже в самих анализаторах (подробнее см. Методы искусственного интеллекта, методы распознавания образов). В результате, основной фокус перемещается на создание [корпусов](http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127206:article), снабженных правильной [лингвистической разметки](http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127221:article) и создание необходимых словарных источников, используемых для обучения. Специалистами в области компьютерной лингвистики создаются интегральные пакеты автоматического анализа текста, включающие всю технологическую цепочку лингвистической обработки. Нередко такие системы усиливаются модулем разрешения анафоры и распознаванием именованных сущностей (имен собственных), например, система Gate, NLTK (natural language toolkit) и др.

Автоматическая обработка звучащей речи и прикладная фонетика

Фонетические (и шире лингвистические) знания, наряду с такими дисциплинами, как цифровая обработка сигналов и математические основания распознавания образов создают теоретическую базу компьютерных [речевых технологий](http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127472:article), значимость которых для современного информационного и компьютеризованного общества трудно переоценить. Главная задача [речевых технологий (РТ)](http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127472:article) состоит в том, чтобы обеспечить удобное и естественное для человека общение с компьютерной техникой. Очевидно, что это должно быть голосовое общение на естественном языке, предполагающее создание средств автоматического ввода/вывода звучащей речи в компьютерные устройства.

Основные задачи автоматического анализа текста

Данное направление лежит на стыке теории информационного поиска, автоматической обработки данных и автоматической обработки естественного языка. Системы и технологии различаются по тому, в какой степени в них представлен собственно лингвистический компонент.

Лингвистическое обеспечение поисково-информационных систем

Одной из прикладных задач обработки текста является информационный поиск. Основные методы информационного поиска разрабатываются в рамках одноименной научной дисциплины. Однако в настоящее время большинство таких систем включают лингвистический компонент - [лингвистическое обеспечение информационных систем](http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127457:article) – способствующий улучшению результатов поиска.

Современные поисково-аналитические системы, как правило, включают себя ряд сопуствующих сервисов, кроме собственно поискового, направленных на систематизацию выдаваемой информации и оптимизацию ее восприятия. Разработкой этих сервисов также занимается компьютерная лингвистика. Это:

1. автоматическая рубрикация документов – тексты разбиваются на группы по темам и жанрам. Например, пользователя, интересующегося новейшими научными разработками в области мобильных телефонов, не интересуют тексты чисто рекламного характера
2. автоматическая классификация и кластеризация текстов: разбиение текстов на группы в соответствие с заранее заданным классификатором (классификация) или же на основании тематически близкого содержания (кластеризация). Часто в качестве ответа на запрос пользователя выдается только несколько наиболее релевантных заголовков или резюме, а пользователем далее может более подробно посмотреть документы, попавшие в один кластер. Особенно активно такие технологии применяются в новостных сервисах, где новости по событиям группируются в “новостные сюжеты”, см, например, [www.news.yandex.ru](http://www.news.yandex.ru/), [www.google.ru](http://www.google.ru/), и т.д.
3. автоматическое реферирование – генерация текста из наиболее значимых предложений документа или группы документов. Активно применяется, например, в новостных сервисах

Задачи извлечения информации из текста, извлечение знаний (Data Mining)

Бурный рост объема текстов, в которых ведется поиск, привел к тому, что те статистические методы, которые сделали возможным быстрый и эффективный поиск по большим массивам неструткурированных данных, стали “мешать” эффективности этого поиска. Преимущества этих методов - отсутствие необходимости подробного семантического описания предметной области и содержательного анализа текстов, породило и его ограниченность.

В последнее время на первый план выходит задача предварительной “когнитивной обработки” текстов. С одной стороны, борьба с “переизбытком” выдаваемой пользователю информации вылилась в то, что современные системы анализа текстов от задачи информационного поиска: найти документ по заданной тематике, переходят к задаче извлечения информации из текстов **(information extraction)** и более глубокого анализа извлеченной информации - извлечения знаний (data mining). В результате пользователь по своему запросу получает не “мешок” текстов, а некоторые “суммирующие” данные, определенным образом структурированные. Если приписать тексту некоторые “семантические” метки, это позволит частично уменьшить количество “шума” - текстов не из той области знаний, которая интересует пользователя, и поможет решить проблему омонимии. Таким образом, на первый план выдвигается задача разработки специальных языков и систем, описывающих понятийную структуру той или иной области знаний - **тезаурусов и онтологий, в том числе и лингвистически-ориентированных онтологий**, создание так называемых универсальных языков. (см. презентацию –семантический модуль)

Новые сервисы, активно использующие лингвистические компоненты анализа (синтаксический и семантический анализ, разрешение анафоры, решают следующие задачи:

1. извлечение из текста именованных сущностей - персон, географических названий, названий организаций и т.п. (а также вычисление "популярности" этой сущности на основе частоты ее упоминания);
2. извлечению фактов (например, автоматическое извлечение из новостных сообщений информации о выступлении политического деятеля, биографии ньюсмейкера ) в СМИ; информации об изменении в финансовом положении некоторой компании и т.п.);
3. извлечение мнений (**opinion mining/extraction, sentiment extraction/analysis, subjuctivity analysis** и т.п., например, для изучения покупательского спроса используется автоматический анализ блогов, в которых обсуждаются те или иные товары или для изучения популярности того или иного политического направления);
4. автоматическое реферирование (например, снабжение аннотацией кластера новостных сообщений).

Системы, моделирующие языковое взаимодействие

Одним из актуальных вопросов компьютерной лингвистики является моделирование языкового взаимодействия, включая языковое взаимодействия компьютера с человеком. Эта задача рассматривалась как одна из задач искусственного интеллекта. В рамках такой задачи был создан целый ряд экспериментальных систем, призванных проверить модели, таких как процедурная модель Винограда, модель концептуальных схем Шенка и др. Первый обзор проблемы синтеза текстов, изданный на русском языке, появился в 1990 году в сборнике “Искусственный интеллект” [5]. В нем кратко перечислена основная проблематика моделирования языкового взаимодействия:

1. теория дискурса,
2. фокус внимания,
3. референция,
4. проблема представления знаний (основными моделями представления в компьютерных системах автоматического понимания ЕЯ, как и во многих других системах искусственного интеллекта являются **фреймы и концептуальные схемы**).

К прикладным системам, включающим модель языкового взаимодействия, относятся такие системы, как:

1. **вопросно-ответные системы**
2. [системы генерации текстов](http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127466:article).

**Ближайшие перспективы прикладной лингвистики**

Современная прикладная лингвистика почти столь же многообразна, как и области практической деятельности человека.  
К наиболее перспективным прикладным её областям в первую очередь следует отнести гипертекстовые технологии, непосредственно связанные с эксплуатацией и развитием глобальной компьютерной сети Интернет. В настоящее время ведётся активная разработка программ-оболочек гипертекста, которые позволяют организовать удобное взаимодействие компьютерной программы с пользователем. При этом достижение, успехов в создании таких программных продуктов невозможно без привлечения фундаментальных знаний о языковой системе.  
В наше время, когда любой уважающий себя Интернет-ресурс имеет страницы, "оснащённые" программами автоматического (машинного) перевода слов и текстов с одного языка на другой, актуальным становится перевод не отдельных слов или словосочетаний, а целых Фрагментов текста. При этом сложности возникают со специальными либо литературными текстами. И одним из направлений совершенствование подобных программ является совершенствование лингвистических алгоритмов, лежащих в основе таких программ.123:27Другая бурно развивающаяся проблематика, связанная с компьютерными технологиями и имеющая существенный лингвистический потенциал, - компьютерный дизайн текста и его компонентов, в том числе шрифта. Имеющееся программное обеспечение позволяет выводить на монитор компьютера всю полосу печатного текста книги, журнала, газеты, рекламного буклета и компоновать текст и изобразительные элементы оригинал-макста в единое целое. В этом случае текст выступает как элемент изображения, а изображение - как часть текста. Чисто практичсские принципы построения оригинал-макета издания должны поддерживаться исследованиями в области лингвистики и семиотики, выявляющими законы сочетания содержания текста, изображения, семантики гарнитуры (шрифта) и способов выделения компонентов текста (курсив, разрядка, отступ и т.д.).  
Имеющийся в лингвистике и семиотике теоретичсский задел - например, разработанная еще в 1920-х годах теория "защит" А.А.  
Реформатского - пока практически не востребована. Однако можно предполагать, что в ближайшее время элементы этой теории будут более широко применяться для решения лингвистических задач в современных информационных технологиях.123:27Особо следует сказать о российской прикладной лингвистике. Для отечественных лингвистов сравнительно недавно открылись те области приложения знаний о языке, которые вполне традиционны для зарубежного сообщества профессиональных лингвистов. В России появился ощутимый спрос на лингвистов, работающих в области рекламы и в сфере публичной политики. Эти два направления объединились в деятельности пиарагентств различного рода. предоставляющих услуги по продвижению на "политическом рынке" как отдельных политиков, так и политических партий, движений. Особенно востребованной является организация предвыборных компаний, связи с общественностью, мониторинг состояния общественного сознания и т.п.  
Здесь следует отметить также широкие возможности использования лингвистических приёмов для соответствующей подачи информации в СМИ с целью формирования общественного мнения по тем или иным вопросам.  
Таким образом, перспективы развития в России лингвистических приложений в рекламе и практической политике весьма значительны.1