Тема: ПРИКЛАДНЫЕ РАЗДЕЛЫ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

План

1. Корпусная лингвистика
1.1 Корпусная лингвистика как раздел прикладной лингвистики.
1.2 Понятие корпуса, разметки. Виды корпусов.
1.3 Требования к корпусам.
2. Компьютерная лексикография
2.1 Понятие компьютерной лексикографии.
2.2 Электронный словарь. Состав словарной статьи. Виды электронных словарей. Преимущества электронных словарей.
2.3 Перспективы компьютерной лексикографии.

1.1 Корпусная лингвистика как раздел прикладной лингвистики

Корпусная лингвистика - раздел прикладной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов при помощи компьютеров.

Корпусная лингвистика включает два аспекта:
1) создание корпусов текстов с автоматическими инструментами их использования;
2) разработка способов экспериментальных исследований различных уровней языка на базе корпусов разных типов.

Корпусы могут использоваться:
1) *в лексикографии* для создания словарей, определения значения многозначных слов и т.д.;
2) *в грамматике* для определения частоты морфем, типов словосочетаний и предложений и т.д.;
3) *в лингвистике* текста для дифференциации типов текста, выявления связей внутри абзаца и между абзацами и т.д.;
4) *в автоматическом переводе* текстов для поиска контекстов слов, имеющих несколько переводных эквивалентов, поиска переводных эквивалентов в параллельных текстах и т.д.;
5) *в учебных целях* для выбора цитат, фрагментов произведений, примеров для организации учебных занятий, создания учебных пособий и т.д.
6) *в тестировании* программ автоматического анализа и синтеза речи и т.д.

1.2 Понятие корпуса, разметки. Виды корпусов.

Лингвистический корпус - совокупность специально отобранных текстов, размеченных по различным лингвистическим параметрам и обеспеченных системой поиска.

Корпус = тексты + их разметка.

Выделяются размеченные (аннотированные) и неразмеченные корпусы текстов. В качестве неразмеченных корпусов можно рассматривать существующие электронные коллекции текстов: виртуальные библиотеки, архивы электронных версий периодических изданий или новостных лент, которые оказываются достаточными для некоторых исследовательских и учебных целей.

Предметом корпусной лингвистики являются преимущественно размеченные корпусы текстов.

Первым этапом в создании корпуса является отбор текстов.

Важным этапом создания корпуса является его разметка.
Разметка (англ. *tagging, annotation)* - это приписывание текстам и их компонентам специальных меток.

Метки могут быть ***внешними*** (экстралингвистическими), включающими сведения об авторе и о тексте, или ***внутренними***: структурными или собственно лингвистическими.

Внешние *метки* содержат сведения об авторе, названии текста, годе и месте издания, жанре, тематике. Сведения об авторе могут включать не только его имя, но также возраст, пол, годы жизни и многое другое. Это кодирование информации имеет название *метаразметка.*
Структурные *метки* несут информацию о статусе каждой единицы (глава, абзац, предложение, словоформа), а ***собственно лингвистические*** описывают лексические, грамматические и прочие характеристики элементов текста.

В соответствии с уровнем лингвистического описания различают *морфологическую* (определение части речи и морфологических категорий), *синтаксическую* (определение синтаксических связей), *семантическую* (категории, характеризующие значение слова), *анафорическую* (характеристика референтных связей, например, местоимений), *просодическую* (характеристика ударения и интонации), *дискурсную* (обозначение пауз, повторов, оговорок устной речи) и некоторые другие виды разметки.

Этой весной опять расцвела акация

Этой - МЖЕТ21 весной - СЖЕТ22 опять - Н22 расцвела - ГЖЕП33 акация - СЖЕИЧ42

Первый индекс указывает на часть речи (М - местоимение, С - существительное, Н - наречие, Г - глагол), второй обозначает род, третий - число, четвертый - падеж или время (у глагола), первая цифра указывает на число слогов, а вторая - на ударный слог.

Виды меток:
<STDS> (стандартное написание), <KOKS> (использование только строчных букв), <KOGS> (использование только прописных букв), <GDOP> (удвоение графем), <GAUS> (выпадение графем), <GZUV> (написание лишней графемы) и т.д.

Таблица **Классификация корпусов**

№	Признак	Виды корпусов
1	Форма хранения	звуковые письменные смешанные
2	Язык текстов	русский английский и т.д.
3	«Параллельность»	одноязычные двуязычные многоязычные
4	Стиль	литературные диалектные разговорные публицистические терминологические смешанные
5	Способ доступа	свободно доступные коммерческие закрытые
6	Разметка	размеченные неразмеченные
7	Характер разметки	морфологические синтаксические семантические просодические и т.д.

Универсальный национальный корпус - это собрание текстов конкретного естественного языка, представительное по отношению ко всему языку, которое может служить для исследования самых разнообразных явлений этого языка. Общепризнанный образец универсального национального корпуса *Британский национальный корпус* (BNC) (www.natcorp.ox.ac.uk). Для русского языка таким представительным корпусом является *Национальный корпус русского языка* (НКРЯ) (www.ruscorpora.ru). Среди корпусов славянских языков выделяется *Чешский национальный корпус* (http://ucnk.ff.cuni.cz), созданный в Карловом университете Праги.

Корпусный менеджер (англ. corpus manager) — это специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме.

1.3 Требования к корпусам.

Требования к созданию корпусов:
1) *репрезентативность* (частота явления в корпусе должна соответствовать его частоте в естественном языке);
2) полнота (явление должно включаться в корпус, даже если его появление не соответствует идее репрезентативности);
3) *достаточный объем* (если первые корпусы достигали миллиона слов, то объем современных корпусов исчисляется сотнями миллионов и миллиардами, например, объем корпуса английского языка *Bank of English* превышает 2,5 млрд слов);
4) *экономичность* (корпус текстов должен экономить усилия исследователя при изучении проблемной области, т.е. быть не просто строгим подмножеством текстов проблемной области, но, по возможности, быть наиболее «экономичным»);
5) *структуризация материала* (в корпусе должны быть выделены адекватные корпусу единицы хранения);
6) *компьютерная поддержка* (поддержка корпуса текстов комплексом программ по обработке данных, обеспечивающих выявление контекстов слова, статистическую инвентаризацию, автоматическую словарную обработку и т.д.).

2.1 Понятие компьютерной лексикографии

Компьютерная лексикография представляет собой раздел прикладной лингвистики, нацеленный на создание компьютерных словарей, лингвистических баз данных и разработку программ поддержки лексикографических работ.

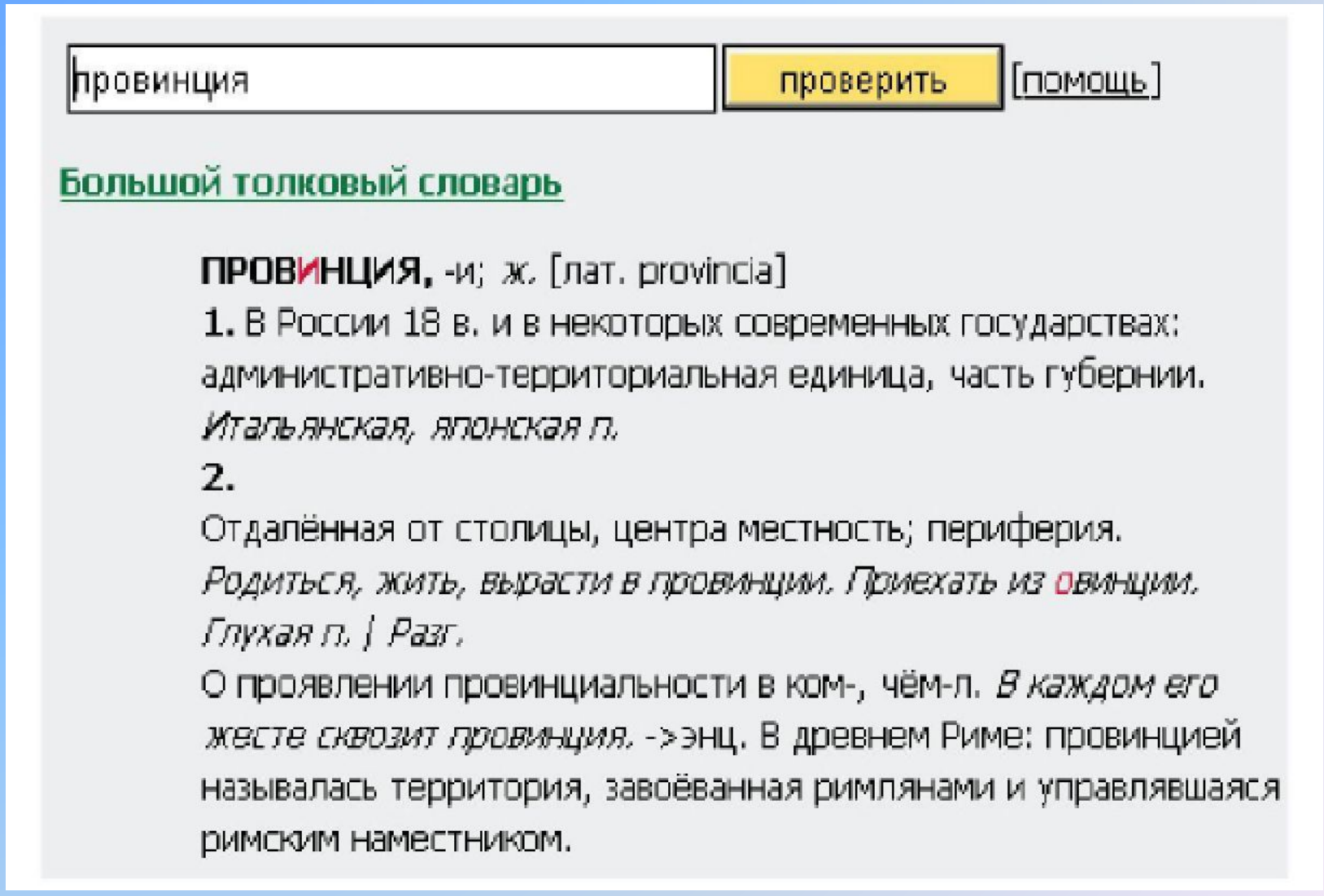
2.2 Электронный словарь. Состав словарной статьи. Виды электронных словарей. Преимущества электронных словарей.

Словарь организованное собрание слов с комментариями, в которых описываются особенности структуры и/или функционирования этих слов. ***Электронный*** (автоматический, компьютерный) ***словарь*** - это собрание слов в специальном компьютерном формате, предназначенное для использования человеком или являющееся составной частью более сложных компьютерных программ (например, систем машинного перевода). Соответственно, различаются *автоматические словари конечного пользователя-человека* (АСКП) и *автоматические словари для программ обработки текста* (АСПОТ).

Компьютерные версии известных словарей:

- Оксфордский словарь английского языка (www.oed.com),
- автоматический толковый словарь английского языка издательства «Коллинз» (www.mycobuild.com),
- автоматический вариант «Нового большого англо-русского словаря» под ред. Ю.Д. Апресяна и Э.М. Медниковой (http://eng-rus.slovaronline. com),
- словарь Ожегова онлайн (http://slovarozhegova.ru).

Рис. Статья компьютерного словаря [Большой толковый словарь: www.gramota.ru]


провинция проверить [помощь]

Большой толковый словарь

ПРОВИНЦИЯ, -и; *ж.* [лат. provincia]

1. В России 18 в. и в некоторых современных государствах: административно-территориальная единица, часть губернии. *Итальянская, японская п.*

2.

Отдалённая от столицы, центра местность; периферия. *Родиться, жить, вырасти в провинции. Приехать из провинции. Глухая п.* | *Разг.*

О проявлении провинциальности в ком-, чём-л. *В каждом его жесте сквозит провинция.* -> энц. В древнем Риме: провинцией называлась территория, завоёванная римлянами и управлявшаяся римским наместником.

Структура *традиционного словаря* обычно включает следующие **компоненты**:

- ✔ введение, объясняющее принципы пользования словарем и дающее информацию о структуре словарной статьи;
- ✔ словник, включающий единицы словаря: морфемы, лексемы, словоформы или словосочетания; каждая такая единица с соответствующим комментарием представляет собой словарную статью;
- ✔ указатели (индексы);
- ✔ список источников;
- ✔ список условных сокращений и алфавит.

Структура словарной статьи включает следующие зоны словарной статьи:

- лексический вход (вокабула, лемма);
- зона грамматической информации;
- зона стилистических помет;
- зона значения;
- зона фразеологизмов;
- зона этимологии;
- зона примера и источника примера.

В *лингвистических словарях* описываются сами слова - их значения, особенности употребления, структурные свойства, сочетаемость, соотношение с лексическими системами других языков и т.д. В *энциклопедических словарях* описываются понятия, факты и реалии окружающего мира, т.е. экстралингвистическая информация. Промежуточный тип словарей включает информацию и лингвистического, и экстралингвистического рода.

Виды лингвистических словарей :

- ✔ *толковые*;
- ✔ *словари-тезаурусы*;
- ✔ *двуязычные (переводные) словари*;
- ✔ *ассоциативные словари*;
- ✔ *исторические и этимологические словари*;
- ✔ *словари языковых форм*;
- ✔ *словари речевого употребления*;
- ✔ *ономастиконы*.

Электронные энциклопедии: Энциклопедия Британника (www.britannica.com), «Большая энциклопедия Кирилла и Мефодия» (www.megabook.ru) и энциклопедия «Кругосвет» (www.krugosvet.ru).

Примеры *переводных электронных словарей:* ABBYY Lingvo (www.lingvo.ru), TranslateIt! (www.translateit.ru) и Multitran (www.multitran.ru).

Электронные толковые словари: словарь Merriam Webster (www.merriam-webster.com) и словарь французского языка «Tresor de la langue francaise» (http://atilf.atilf.fr).

Формальныме электронные словари: орфографические словари русского (http://slovari.yandex.ru) и английского (www.spellcheckonline.com) языков.

Большую коллекцию словарей разных видов на дисках и в Интернете предоставляет издательство *Duden* (немецкий язык, www.duden.de) и *Larousse* (французский язык, www.larousse.fr).

Создание электронного словаря включает следующие **этапы**:

1) формирование корпуса текстов и параллельно создание словника;
2) автоматическое формирование корпуса примеров;
3) написание словарных статей;
4) ввод словарных статей в базу данных (БД);
5) редактирование словарных статей в БД;
6) корректура текста в БД;
7) порождение текста словаря и формирование оригинал-макета;
8) печать словаря.

2.3 Перспективы компьютерной лексикографии.

Для лексикографических целей - базы данных типа ACCESS или PARADOX.

Для поиска примеров - компьютерные программы построения конкордансов, например, DIALEX.

Для создания оригинал-макета (верстки) словарей привлекаются издательские системы типа Page-Maker или WinWord

Пример специализированной компьютерной программы, предназначенной для компьютерных лексикографических работ - «Программа автоматизированного составления и обработки словников» (авторы: М.В. Литус, Е.В. Литус).

Преимущества в использовании электронных словарей:
1) электронные словари позволяют по-разному представить содержание словарной статьи (различные «проекции» словаря), в том числе с помощью разнообразных графических и мультимедийных средств, которые не используются в обычных словарях;
2) в выдаваемой информации находят отражение различные технологии компьютерной лингвистики, например морфологический и синтаксический анализ, полнотекстовый поиск, распознавание и синтез звука и т.п.;
3) становится возможным быстро получить информацию, которая содержится где-то в недрах словаря и непосредственно отвечает тому запросу, который сформулирован пользователем в удобной для него форме;
4) электронный словарь позволяет быстро реагировать на изменения в языке и мире, и выпуск каждой последующей его версии или внесение изменений в онлайн-версию не занимает много времени и труда.

Нерешенные проблемы, актуальные как для традиционной, так и для компьютерной лексикографии.

1. В словарях должно найти отражение понятие *лексической функции,* позволяющее систематически описывать несвободную сочетаемость слов, иллюстрируемую следующими примерами русского языка: «войну ведут», а «экзамен - держат», «теории выдвигают», а «мысли подают» и т.п.
2. Не нашла отражение в массовой лексикографической практике проблема *описания семантики* и практической *реализации грамматического словоизменения* и *словообразования.* Каждый язык имеет свои собственные способы грамматического кодирования смысла, которые не описываются в массовых словарях систематически.
3. Например, как передать по-английски смысл «довыпендриваться», даже если знаешь, как передать «выпендриваться»?
4. В словарях не существует даже системы понятий, с помощью которой *синтаксическая информация* могла бы быть доведена до обычного читателя. Решением этой проблемы могли бы стать интегральные словарные описания, основанные на формальных моделях, учитывающие прогрессивные лексикографические идеи. На этих же моделях следует организовать технологии доступа к словарному содержанию.

СПАСИБО ЗА ВНИМАНИЕ!