Компютърна компаративистика

Езикови технологии за славистични изследвания

SlavicNLP@online.de

ст.н.с. II ст. д-р **Елена Паскалева**, ИПОИ БАН

PD Dr. phil. habil. **Таня Августинова**, DFKI & Saarland University

Компютризация на славистичните изследвания

- (структурна лингвистика)
- приложна и математическа лингвистика
- компютърна лингвистика
  - възникване на нови информационни области
  - информатизация на филологическото образование
- формализация на текста в информационните технологии
  - ортографически и граматически коректори
  - системи за разпознаване на печатни и ръкописни текстове
  - системи за разпознаване на реч
  - системи за търсене и извличане на информация
  - системи за автоматичен превод и електронни речници
  - системи за автоматично резюмиране на документи
  - системи за обектно-графичен анализ на текста
- лингвистични основи на информатиката
  - формално-граматично описание на езиковата структура
  - съпоставително изследване на граматическите формализми
  - специализирана обработка на различните нива на езиково описание
  - модулен подход при създаване на компютърни граматики

Пример за тенденции в съвременната русистика

(основните теми на конференцията МАПРЯЛ 2007)

1. *Новое в системно-структурном описании современного русского языка*
   - Типологические особенности русского языка.
   - Структурный, семантический, функциональный аспекты изучения языковых единиц разных уровней.
   - Фонетика и фонология.
   - Морфемика и морфология.
   - Словообразование.
   - Лексикология.
   - Фразеология.
   - Синтаксис.
   - Лингвистика текста и дискурс.
2. *Речевая деятельность: современные аспекты исследования*
   - Коммуникативно-прагматические аспекты изучения единиц языка.
   - Стуктура речевого акта.
   - Типы речевых актов и речевых действий.
   - Интенциональная структура высказывания.
   - Речевые регистры.
   - Коммуникативное поведение.
   - Речевые стратегии и тактики: режимы согласия и конфликта, языковая манипуляция, языковая демагогия, языковая игра.

Пример за тенденции в съвременната русистика
(основните теми на конференцията МАПРЯЛ 2007)

3. *Функциональные разновидности русского языка*
   - Сферы общения и функциональные стили языка и речи.
   - Языки различных профессиональных и социальных групп: русский язык СМИ, бизнес-язык, Интернет-общение, языки русской диаспоры, гендерная стилистика.
   - Язык художественной литературы.
4. *Язык, сознание, культура*
   - Когнитивная интерпретация русских языковых фактов.
   - Фрагменты русской концептосферы.
   - Дисциплинарный статус лингвокультурологии.
   - Методология и методы лингвокультурологии.
   - Базовые понятия лингвокультурологии.
   - Концептуальная и языковая картина мира.
   - Этническая ментальность.
   - Русская языковая личность.
   - Национально-культурная маркированность языковых единиц.
   - Концептосфера русского речевого общения.

Пример за тенденции в съвременната русистика
(основните теми на конференцията МАПРЯЛ 2007)

5. *Русский язык: диахрония и динамика языковых процессов*
   - Развитие русского языка на протяжении его истории.
   - Язык древнерусской книжности.
   - Электронная обработка древнерусских рукописей.
   - Становление и эволюция норм литературного языка.
   - Социокультурная и социолингвистическая проблематика литературной нормы.
   - Актуальные процессы в русском языке конца XX – начала XXI столетия.
   - История русской лингвистической мысли.
6. *Русская лексикография: тенденции развития*
   - Лексикографическое представление русского языка в словарях разных типов.
   - Теория и практика лексикографического описания.
   - История лексикографии.
   - Терминоведение и терминография.
   - Учебная лексикография.
   - Словарь и межкультурная коммуникация.
   - Корпусы текстов и лексикография.
   - Словарь и Интернет.
   - Электронная лексикография.
   - Проекты словарей нового поколения.
   - Презентация электронных корпусов и словарей.

Пример за тенденции в съвременната русистика
(основните теми на конференцията МАПРЯЛ 2007)

7. *Русский язык в сопоставлении с другими языками*
   - Методология межъязыковых сопоставлений.
   - Проблемы таксономической и объяснительной типологии.
   - Проблемы межъязыковой эквивалентности.
   - Системная, функциональная и прагматическая эквивалентность.
   - Универсальное и идиоэтническое в русском языке.
   - Способы языковой кодировки в русском и сопоставляемых с ним языках.
   - Прикладные аспекты описания русского языка в сопоставлении с другими языками.
   - Вопросы составления корпусов параллельных текстов.
8. *Коммуникация на русском языке в межкультурной среде*
   - Этнопсихолингвистические, социокультурные и прагматические аспекты межкультурной коммуникации.
   - Межкультурная коммуникация и дискурсивные стратегии.
   - Невербальные компоненты коммуникации на русском языке в условиях межкультурных контактов.
   - Бизнес-коммуникация на русском языке в межкультурной среде.
   - Интернет как особая коммуникативная среда.
9. *Перевод – взаимодействие языков и культур*
   - Современная теория и практика перевода.
   - Функциональное, содержательное и структурное соотношение перевода и оригинала.
   - Универсальное, национальное и индивидуальное в тексте перевода.
   - Проблема интерпретации при переводе.
   - Перевод в диалоге культур.
   - Перевод в сфере профессиональной коммуникации.
   - Компьютеризация перевода.

Пример за тенденции в съвременната русистика

(основните теми на конференцията МАПРЯЛ 2007)

10. Изучение и описание русского языка как иностранного

- Теоретические и прикладные модели описания русского языка как иностранного.
- Специфика лингвистической модели русского языка как неродного.
- Дифференциация лингвистического описания русского языка как неродного и как иностранного для всех уровней обучения.

11. Методика преподавания русского языка (родного, неродного, иностранного)

- Актуальные проблемы школьного, вузовского и курсового преподавания русского языка.
- Язык и культура, проблематика межкультурной коммуникации в теории и практике преподавания русского языка.
- Проблемы преподавания русского языка в полиэтническом обществе.
- Теория и практика национально ориентированного обучения.
- Программы, учебники, учебно-методические пособия в обучении русскому языку.
- Современные технологии обучения: мультимедийные обучающие программы, программы для компьютерной диагностики и коррекции языковых ошибок…
- Дистанционное обучение.
- Теория и практика тестирования уровней владения русским языком.
- Повышение квалификации преподавателей.

12. Русская литература: история и современность

13. Методика преподавания русской литературы

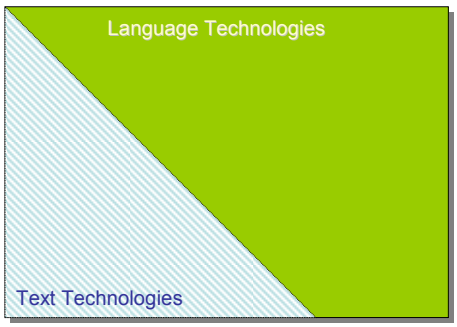
Лингвистично моделиране

- търсене и клсификация (http://www.dialog-21.ru/trends/?id=1772)
  - словоизменението като основна задача на приложната морфология
  - проблеми на словообразуването при автоматично търсене на информация
- роля на формалните граматики (http://www.dialog-21.ru/trends/?id=2026)
  - синтактични модели за генериране и разпознаване на текстовата структура
  - структура на изречението в системите за обработка на текст
  - представяне на синтактичната структура: депендентна граматика, конституентна граматика
  - формални свойства на синтактичната структура на изречението
- компютърна лексикография (http://www.dialog-21.ru/trends/?id=2024)
  - автоматизация на лексикографската дейност
  - основни приципи за изграждане на компютърен речник
  - програми за създаване и поддръжка на речници
  - бази от данни, електронни картотеки, текстообработващи програми
- корпусна лингвистика (http://www.dialog-21.ru/trends/?id=1278)
  - качествено нова парадигма в езиковедските изследвания
  - корпусни технологии, програмно осигуряване при работа с корпуси
  - типове корпуси, анотационни схеми, международни стандарти
- машинен превод (http://www.dialog-21.ru/trends/?id=1744)

Езикови технологии

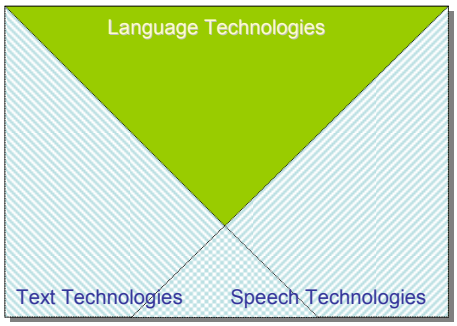
[Quelle: Computerlinguistik UdS]

Езикови технологии



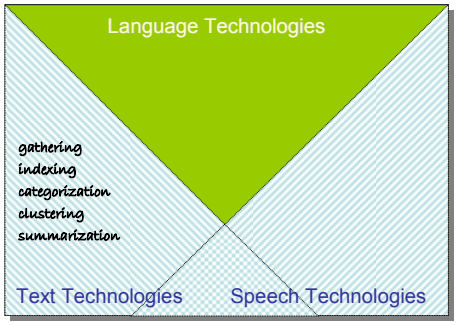
[Quelle: Computerlinguistik UdS]

Езикови технологии



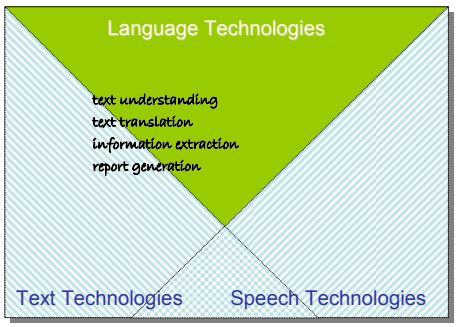
[Quelle: Computerlinguistik UdS]

Езикови технологии



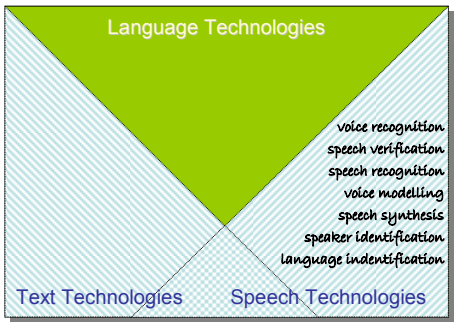
[Quelle: Computerlinguistik UdS]

Езикови технологии



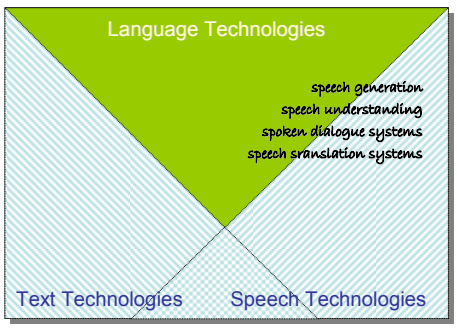
[Quelle: Computerlinguistik UdS]

Езикови технологии



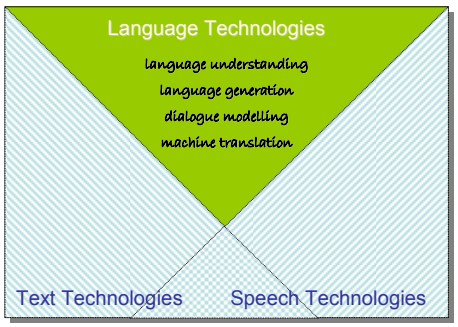
[Quelle: Computerlinguistik UdS]

Езикови технологии



[Quelle: Computerlinguistik UdS]

Езикови технологии



[Quelle: Computerlinguistik UdS]

Speech Recognition

- Spoken language is recognized and transformed into text as in dictation systems, into commands as in robot control systems, or into some other internal representation.

[Quelle: Computerlinguistik UdS]

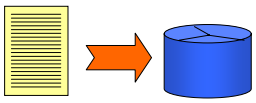
Speech Synthesis

- (also Speech Generation)
- Utterances in spoken language are produced from text (text-to-speech systems) or from internal representations of words or sentences (concept-to-speech systems).

[Quelle: Computerlinguistik UdS]

Text Categorisation

- This technology assigns texts to categories.
- Texts may belong to more than one category, categories may contain other categories.
- Filtering is a special case of categorization with just two categories.



[Quelle: Computerlinguistik UdS]

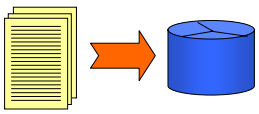
Text Summarisation

- The most relevant portions of a text are extracted as a summary.
- The task depends on the needed lengths of the summaries.
- Summarization is harder if the summary has to be specific to a certain query.

[Quelle: Computerlinguistik UdS]

Text Indexing

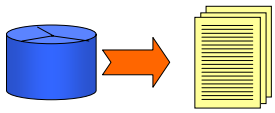
- As a precondition for document retrieval, texts are stored in an indexed database.
- Usually a text is indexed for all word forms or, after lemmatization, for all lemmas.
- Sometimes indexing is combined with categorization and summarization.



[Quelle: Computerlinguistik UdS]

Text Retrieval

- From a database, texts are retrieved that best match a given query or document.
- The candidate documents are ordered with respect to their expected relevance.
- Indexing, categorization, summarization and retrieval are often subsumed under the term information retrieval.



[Quelle: Computerlinguistik UdS]

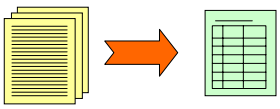
Information Extraction

- Relevant pieces of information are discovered and marked for extraction.
- The extracted pieces can be:
  - the topic, named entities such as company, place or person names,
  - simple relations such as prices, destinations, functions etc.
  - or complex relations describing accidents, company mergers or football matches.

[Quelle: Computerlinguistik UdS]

Data Fusion and Text Data Mining

- Extracted pieces of information from several sources are combined in one database.
- Previously undetected relationships may be discovered.



[Quelle: Computerlinguistik UdS]

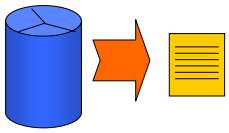
Question Answering

- Natural language queries are used to access information in a database.
- The database may be a base of structured data or a repository of digital texts in which certain parts have been marked as potential answers.

[Quelle: Computerlinguistik UdS]

Report Generation

- A report in natural language is produced to describe the essential contents or changes of a database;
- This report contains accumulated numbers, maxima, minima and the most drastic changes.



[Quelle: Computerlinguistik UdS]

Spoken Dialogue Systems

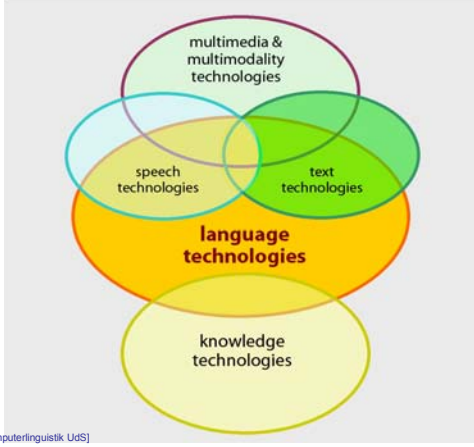
- The system can carry out a dialogue with a human user in which the user can solicit information or conduct purchases, reservations or other transactions.

[Quelle: Computerlinguistik UdS]

Translation Technologies

- Technologies that translate texts or assist human translators.
- Automatic translation is called machine translation.
- Translation memories use large amounts of texts together with existing translations for efficient look-up of possible translations for words, phrases and sentences.

[Quelle: Computerlinguistik UdS]



[Quelle: Computerlinguistik UdS]

Maturity of Speech Technologies

- Voice Control Systems
- Dictation Systems
- Text-to-Speech Systems
- Machine Initiative Spoken Dialogue Systems
- Identification and Verification Systems
- Spoken Information Access
- Mixed Initiative Spoken Dialogue Systems
- Speech Translation Systems

Deployed. On the market
Mature or close to maturity
Research prototypes in R&D

[Quelle: Computerlinguistik UdS]

Maturity of Text Technologies

- Spell Checkers
- Machine-Assisted Human Translation
- Translation Memories
- Indicative Machine Translation
- Grammar Checkers
- Information Extraction
- Human Assisted Machine Translation
- Report Generation
- High Quality Text Translation
- Text Generation Systems

Deployed. On the market
Mature or close to maturity
Research prototypes in R&D

[Quelle: Computerlinguistik UdS]

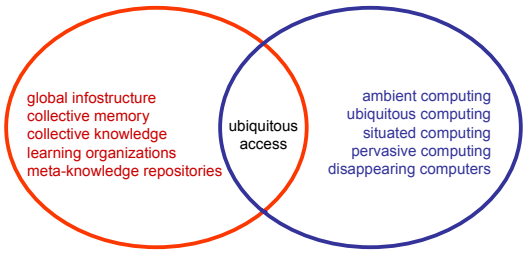
Maturity of Information Technologies

- Word-Based Information Retrieval
- Summarization by Simple Condensation
- Simple Statistical Categorization
- Simple Automatic Hyperlinking
- Cross-Lingual Information Retrieval
- Automatic Hyperlinking With Disambiguation
- Simple Information Extraction (Unary, Binary Relations)
- Complex Information Extraction (Ternary+ Relations)
- Dense Associative Hyperlinking
- Concept-Based Information Retrieval
- Text Understanding

Deployed. On the market
Mature or close to maturity
Research prototypes in R&D

[Quelle: Computerlinguistik UdS]

Mega Trends



personalization
adaptation
learning

[Quelle: Computerlinguistik UdS]