Text Mining in historical text books

A new project on information behavior in the humanities
Mandl: Text Mining in historical text books
Text Mining

- "Discover useful and previously unknown ‘gems’ of information in large text collections" (Avaquest 2002)
- "Text Mining is the extraction of knowledge out of many texts. ... The newly created knowledge could not be read from one single text but only by considering patterns over many texts ... " (Mandl 2013 @ KSS)
- "Process of deriving high quality information from text" (Choi 2013 @ LWA)
Text Mining

Authors → Text → Representation → Result

Creation → Extraction → Surrogat → New Knowledge

Algorithm:
- Machine Learning
- Visualisation
- Temporal Analysis
- ...

Mandl: Text Mining in historical text books
Die Welt der Kinder

Mandl: Text Mining in historical text books
Die Welt der Kinder

The World of the Children

Knowledge about the world and meaning of world in children's text books and children novels between 1850 und 1918
Corpora

Text books (History and Geography)

Children's and teens' books

Knowledge from school books

Popular Knowledge

Knowledge about the world

Mandl: Text Mining in historical text books
Corpora

**GEI-Digital**
- All German history text books of the „Kaiserreich“
- Some 1,800 books (350,000 pages)
- OCR processed
- Metadaten in METS/MODS-Structure
- Digitalisation of Geography books ongoing
- Until 2014 some 3,000 books more will be available (500,000 pages)
- TEI-Format
Corpora

- Youth literature from UB Braunschweig
  - 2500 books,
  - Digitalisation ongoing
  - Genres: Adventure, Robinsonades, songbooks, moral and religious tales
Text Books

- Text book in 19th century as a primary tool to access world knowledge
- Children were obliged to visit school
  - Between 1871 and 1918: 836 new history text books were published
  - Between 1890 and 1918: as many books were published as between 1700 and 1850
Areas of Transmitted Knowledge

„Foreign“ Continents, Countires and Peoples, Colonies, Discoveries, Expeditions,

War and Peace, political treaties, Regents and states men

Science and Culture, Urbanisation and Industrialisation, Pauperism, Social politics and family

Transport, Communication, Trade, Industry, Technology, environment and nature

Mandl: Text Mining in historical text books
Topics

Mandl: Text Mining in historical text books
Goals

● Semantic, categorial and symbolic knowledge about the world
● How did the authorities interpret the world and transfer this knowledge to the children
● Diachronic perspective
● Differences between parts of the corpora
Topics

Krieg und Frieden, politische Bündnisse, Regenten und Staatsmänner

Wissenschaft und Kultur, Urbanisierung und Industrialisierung, Pauperismus, Sozialpolitik und Familie

Transport, Kommunikation, Handel, Industrie, Technik, Umwelt und Natur

Mandl: Text Mining in historical text books
Areas

Krieg und Frieden, politische Bündnisse, Regenten und Staatsmänner

Mandl: Text Mining in historical text books
Untersuchungsfragen

Weltwissen, Weltsichter, Weltdeutung

Perspektiven und Vergleiche:

synchron | diachron | intermedial | transmedial

Selbst- und Fremdbilder, Fortschritts-, Krisen- und Bewältigungsdiskurse, …

Mandl: Text Mining in historical text books
Mandl: Text Mining in historical text books
Methods, Tools, Technologies

Hermeneutics (intellectual text analysis) and qualitative Content Analysis

and

Tools for digital Analysis

• Opinion Mining and Topic Extraction from historic text corpora
• Analysis and extraction of opinionated text parts
• Temporal Development of Topics and Opinions
• Adaptation of Tools to the specific demands of the historians and the structure of the Texts („domain specific text mining“)
• User oriented creation, evaluation and refinement of Software tools
Tasks and Work Packages

- GEI
  - Historical Research
- DIPF
  - Software development
- U Hildesheim
  - Information Behavior Analysis
  - Requirements Specification
  - Evaluation
Wilson's process model of Ellis' behavioral framework of information seeking (adapted from Wilson 1999)
Thank you for your attention