

Exploring Historical Sources with Language Technology: Results and Perspective

Carsten Schnober



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Welt der Kinder Children and their World

<http://welt-der-kinder.gei.de/>

<https://www.ukp.tu-darmstadt.de/research/current-projects/welt-der-kinder/>

Welt der Kinder: People and Institutions



- Funded by Leibniz Association
- May 2014 – May 2017
- Georg Eckert Institute for International Textbook Research
 - Maik Fiedler, Andreas Weiß, Dr. Robert Strötgen, Prof. Dr. Simone Lässig
- Ubiquitous Knowledge Processing Lab (UKP) TU Darmstadt, and German Institute for International Educational Research (DIPF)
 - Carsten Schnober, Dr. Richard Eckart de Castilho, Prof. Dr. Iryna Gurevych
- Institute for Information Science and Language Technology in Hildesheim (IWIST)
 - Ben Heuwing, Prof. Dr. Thomas Mandl, Prof. Dr. Christa Womser-Hacker



Welt der Kinder – Children and their World



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- Background
- Project Goals
- Research Questions
- Methods
 - Natural Language Processing
 - User-centered development
 - Visualization

Welt der Kinder: Background

- Digital humanities: cooperation between historians, information scientists, and computer scientists
- Research subject: German textbooks from 1850 until 1918
 - 6,000 digitized textbooks and juvenile books (900,000 pages)
 - available at GEI Digital: <http://gei-digital.gei.de>
- Period from 1850 till 1918 of particular historical interest
 - Accelerated production of knowledge
 - Dominated by globalization and nationalization simultaneously
- Sources reflect contemporary world interpretation patterns and elements of cultural memory at the time

- 1) Historical research about representations and interpretations of the world
 - Textbooks as main information source for young adults
- 2) Exploration of specific media types: textbooks and juvenile literature
 - Consider media type, coverage, and transformation of collective knowledge
- 3) Combine established hermeneutic methodology with innovative methods
- 4) Fundamental research in computational linguistics
 - Develop and adapt methods of topic modeling, semantic analysis, and opinion mining

Welt der Kinder: Research Questions



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- 1) “Foreign” continents, countries, peoples, colonies, discoveries, expeditions, missions
- 2) War and peace, political alliances, regents and statesmen
- 3) Science and culture, urbanisation and industrialisation, pauperism and social policies
- 4) Transportation, communication, trade, industry, technology, environment



- OCR done mostly with Abbyy Finereader in a previous project
 - Estimated accuracy rate: 90-95% (characters)
- OCR error analysis
 - Segmentation: “DieNordarmeestand unter”, “Hilfs-truppen”
 - Character recognition: “alle feine Lander”, “den Â©rund zu”, “Dtto”
 - Garbage: "ijii1aAll"
- Spelling variations in the original document (unrelated to OCR)
 - Differences to today's language and within the corpus (canonical German orthography was introduced in 1905)
 - Spelling mistakes by authors

Topic Modelling: Sample Topics

- **Words beginning with capital letters (200 topics, top 10 words):**
 - Sieg, Eroberung, Zug, Tod, Herrschaft, Krieg, Iii, Kampf, Unterwerfung, Aufstand
 - Alexander, Perser, Babylon, Cyrus, Syrien, Israel, Euphrat, Aegypten, Asien, Darius
 - Reich, deutsch, Reiche, deutsche, Grenze, groß, Mark, Nation, Herrscher, Nachfolger
- **Automatically detected named entities (200 topics, top 10 words):**
 - Athen, Athener, Sparta, Athens, Griechenland, griechischen, Salamis, athenischen, Xerxes, Aristides
 - Joseph, Isaak, Abraham, Juden, Jakob, Gott, Kanaan, Israel, Abrahams, Gottes
 - Westfalen, Bielefeld, Minden, Paderborn, Dortmund, Herford, Arnsberg, Soest, Hagen, Hamm
- **Sentiment Words (50 topics, top 10 words):**
 - Recht, Schuld, hart, streng, Klage, verbieten, bestrafen, schwer, Weise, schuldig
 - Tod, Erbe, Besitz, Anspruch, Streit, Macht, Recht, anerkennen, Vertrag, Konstanz
 - führen, Macht, erklären, mächtig, unterstützen, unabhängig, vollständig, glänzend, Unterstützung, kräftig

Topic Modelling in *Welt der Kinder*



- Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003)
- Model variants
 - Number of topics: 20, 50, 100, 150, 200, 1000
 - Lemmatization, stopword filtering, ...
 - Word filtering
 - capitalized words (~nouns in German orthography)
 - named entities
 - sentiment words

Identifying Requirements: User-Centered Development

- Interviews and contextual inquiry with historians about previous projects
- Scenarios and wireframes as “boundary objects”
 - Virtual users from observations of real-world historic research processes
- Models and tools continuously adapted to emerging requirements
- Specificity vs. Sustainability
 - Approaches and technologies developed for project requirements, but as generical as possible

Topic Modelling: Identified Requirements



- Topics have to be homogeneous and easy to label
- Comparisons between topics, instead of analysis of terms within topics
- Topic correlations are relevant
 - Correlations between topics, co-occurrences of distinct topics
 - Examples: France ~ Trade? France ~ War? France ~ Peace?
 - Differences in correlations of topics within sub-collections
- Only a small part of the topics is helpful for research
- Macro-analysis with exemplary documents

Topic Modeling and Visualization

- Identify *relevant* topics for document / collection / task
- Give (visual) impression about the distribution of topics
 - Compare sub-collections by time, school type (confession, gender, level), region, ...
- Approaches:
 - Display document-topic distribution over time in the current facet
 - DiTop (Oelke et al., 2014)
 - Serendip (Alexander et al., 2014)



> [PublicationYear:\[1850 TO 1920\]](#)

Topic Model:

Topic:

Topic: T80: [Rom, Römer, Titus, Pyrrhus, Italien, Samniter, Nero, Tarent, Trajan, Hadrian]

▼ Erscheinungsjahr

1850 : 1920

[Von 1850 bis 1920 \(512308\)](#)

[1850 - 1860 \(17245\)](#)

[1860 - 1870 \(24203\)](#)

[1870 - 1880 \(61313\)](#)

[1880 - 1890 \(76544\)](#)

[1890 - 1900 \(86239\)](#)

[1900 - 1910 \(109144\)](#)

[1910 - 1920 \(137620\)](#)

[Alle](#)

▼ Sammlungen

[Kaiserreich](#)

[Geschichtsschulbuecher](#)

(333939)

[Geographieschulbuecher](#)

[Kaiserreich \(133711\)](#)

[Geschichtsschulbuecher vor](#)

[1871 \(33020\)](#)

[Geographieschulbuecher vor](#)

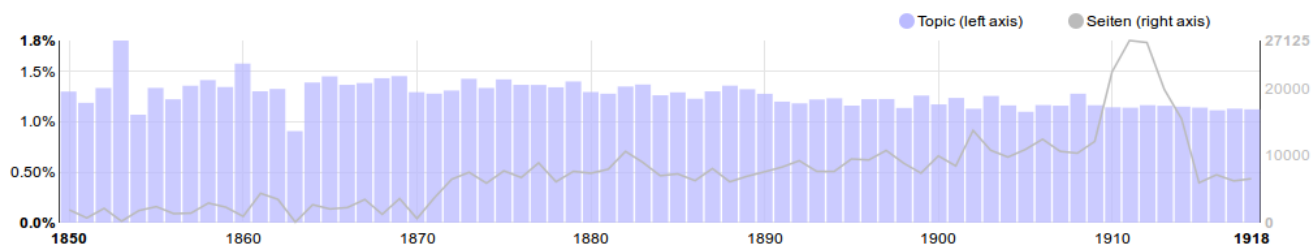
[1871 \(7932\)](#)

[Geschichtsatlanten \(1752\)](#)

[Geographieatlanten \(1360\)](#)

[Realienbuecher vor 1871](#)

(594)



512308 Ergebnisse

Seite 1 von 51231 [weiter >](#)

1. Erzählungen aus der vaterländischen Geschichte - S. 9

Topic 80: 19%

1884 - Leipzig : Siegmund & Volkening

von H. Damm

Sammlung: [Kaiserreich Geschichtsschulbuecher]

Topic(s): [T147: [Tempel, Priester, Spiel, Aegypten, Pyramide, Nil, Fest, Grab, Theben, tote], T29: [Reich, deutsch, Reiche, deutsche, Grenze, groß, Mark, Nation, Herrscher, Nachfolger], T58: [Sieg, Eroberung, Zug, Tod, Herrschaft, Krieg, Ili, Kampf, Unterwerfung, Aufstand], T53: [Auflage, Leipzig, Verlag, Buch, Ausgabe, Preis, Aufl, Unterricht, Karte, Schule], T161: [Alexander, Perser, Babylon, Cyrus, Syrien, Israel, Euphrat, Aegypten, Asien, Darius]]

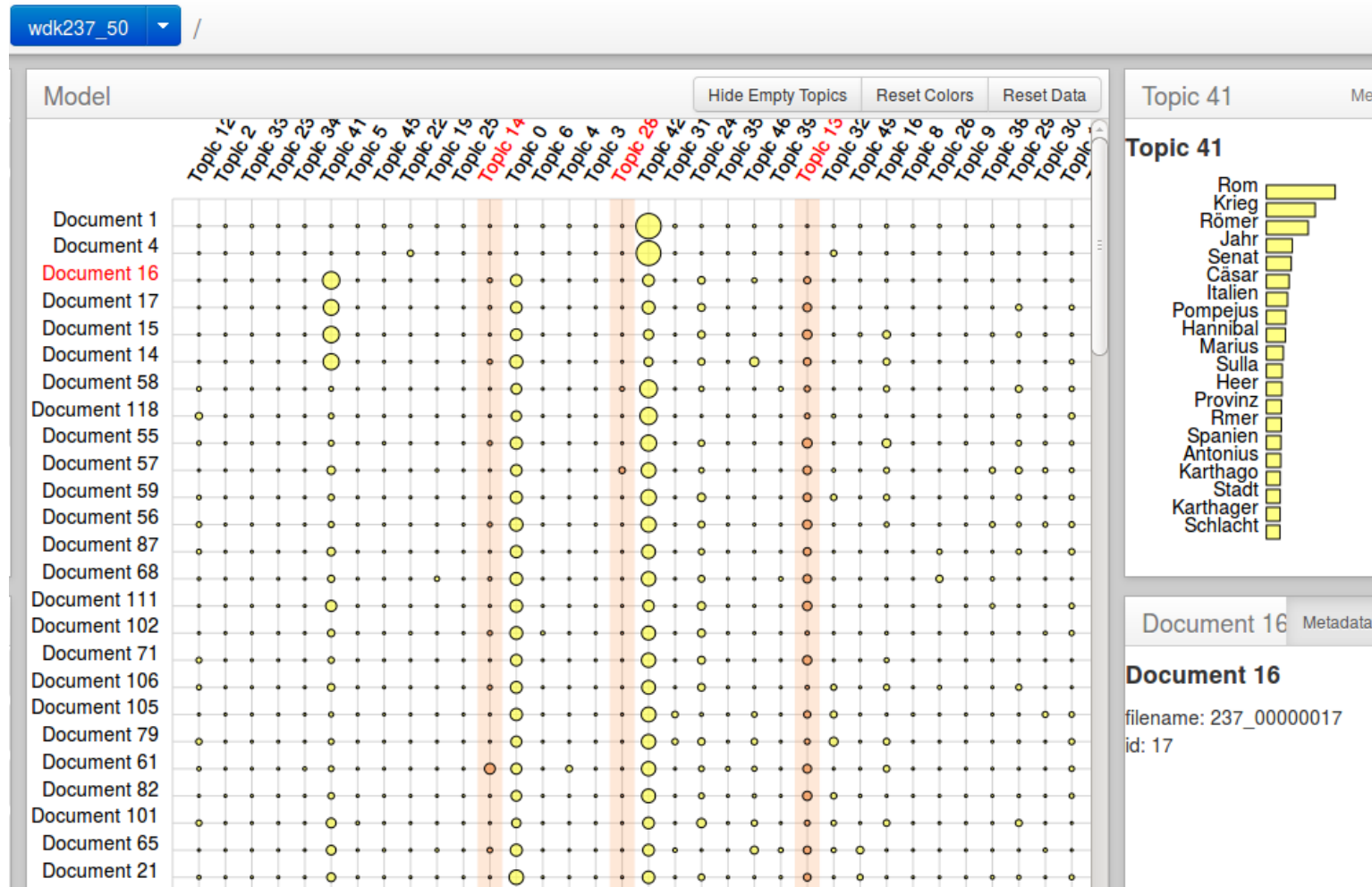
Die Kimbern und Teutonen. — Weitere Kämpfe mit den Römern. 7 ihre Wohnsitze. Auch traten sie zu größeren Vereinigungen zusammen, so daß sich aus ihnen größere Völkerstämme entwickelten. Alle deutschen Stämme aber zeichneten sich durch eine große Freiheitsliebe aus, und daher kam es, daß sie mit den Römern, welche alle andern Völker unterjochen wollten, bald in Kampf gerieten. 2. Die Cimbern und Teutonen. 1. Etwa ums Jahr 113 v. Chr. verließen diese deutschen Völkerstämme ihre Wohnsitze an der Nord- und Ostsee und erschienen unerwartet an der Grenze des römischen Reichs. Dreimalhunderttausend Riesen wären es, so erzählte man sich in Rom, mit blauen Augen, gelblichen Haaren und in niegesehenem Kriegsschmuck; durch ganz Italien verbreitete sich der größte Schrecken. Sie durchzogen die Schweiz und fielen verheerend in Gallien ein. Alle gegen sie ausgesandten römischen Heere würden geschlagen. Jetzt näherten sie sich in zwei getrennten Haufen Italien. Man glaubte nicht anders, als sie würden nun gegen Rom selbst heranziehen und demselben den Untergang bereiten. In dieser Not wählten die Römer zum Konsul. Dieser zog zunächst den Teutonen mit einem Heere entgegen und traf sie ein bei der Rhone, wo er ihnen gegenüber ein verschanztes Lager aufschlug. Nachdem er seine Soldaten in kleinen Gefechten erst an den Anblick, das Kriegsgeheul und an die Fechtweise der Deutschen gewöhnt hatte, griff er sie bei Aquä Sextiä, dem jetzigen Aix im südlichen Frankreich an und vernichtete sie in einer

- Scenario: users interested in certain topics, exploring and searching
- Add documents and topic proportions to Solr index
- Solritas web interface
 - Text search
 - Topic proportion range queries:
`topic1:[0.5 TO *] topic2:[0.0 TO 0.1]`
 - Most relevant topics shown for each document
 - Aggregation / intersection of document selection through facets
 - Documents associated with topic X from model A
 - Documents from publisher P , period Y , ...

Serendip



- Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. "Serendip: Topic Model-Driven Visual Exploration of Text Corpora". In *Proceedings of the 2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* — November 2014



Opinion Mining: *Welt der Kinder* Examples



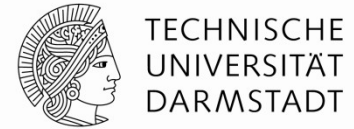
- Obvious opinions
 - Luther war der größte Mann seiner Zeit.
“Luther was the greatest man of his time.”
 - der schreckliche Bauernkrieg
“the terrible Peasant War”
 - hochmütige Franzosen
„arrogant Frenchmen”
- Subtle / complex opinion (requiring world knowledge)
 - Und so förderte die Bibelübersetzung das Reformationswerk unermesslich.
“So the translation of the bible immensely promoted the reformation work.”
 - Japan, das asiatische England
“Japan, the Asian England”

- Little annotated data available in German in general
- Even less (no) data for historical data
- Weakly supervised approaches
 - “Joint Sentiment/Topic Model for Sentiment Analysis” (Lin and He 2009)
 - “Thumbs up or Thumbs Down?”, (Turney, 2002)
 - “Sentiment Analysis on the People’s Daily.” (Li and Hovy, 2014)
 - In a single document (news article), sentiment towards an entity is consistent
 - Over a certain period of time, sentiments towards an entity are inter-related

Summary

- Welt der Kinder
 - Research on historic German textbooks
 - Digitized corpus of 6,000 textbooks and juvenile books
 - Combine traditional methods and natural language processing technologies
- Natural Language Processing
 - OCR (post-processing)
 - Topic modelling
 - Opinion mining
 - Named Entity Recognition
 - ...

Thanks



- Georg Eckert Institute for International Textbook Research
- Ubiquitous Knowledge Processing Lab (UKP) TU Darmstadt, and German Institute for International Educational Research (DIPF)
- Institute for Information Science and Language Technology in Hildesheim (IWIST)
- Institute of Popular Culture Studies, University of Zurich
- Bavarian State Library
- Göttingen Centre for Digital Humanities
- Universitätsbibliothek Technische Universität Braunschweig

