The billion words library - A new implementation of CTS

Gerhard Heyer
Jochen Tiepmar, Christoph Teichmann
Universität Leipzig
heyer@informatik.uni-leipzig.de

UNIVERSITÄT LEIPZIG
Institut für Informatik

Europa fördert Sachsen.
ESF
Europäischer Sozialfonds

ASV
Automatische Sprachverarbeitung

Automatische Sprachverarbeitung

Short Retrospect: Computer Science and its applications

1940-1960	Scientific Computing
70s	Databases, Buisness Computing
80s	Digitizing Electrical Engineering, Beginnings of Text Processing, SGML
90s	Digitizing analogue media, Connecting Distributed Ressources: http, HTML, XML
since 2000	Internet Based Services, Knowledge Managment, Semantic Web

Automatische Sprachverarbeitung

Outline

- Analogue processes and media move to digital ones
- Analog workflows start to incorporate digital aspects
- New possibilities using digital data and algorithms
- Standards - techniques and interfaces - as „enabler“ (HTTP and TCP/IP, HTML/XML)

Moving to digital media and processes creates new workflows and applications and enables Linking

Automatische Sprachverarbeitung

Outline

- **Motivation - The need for a Billion Word Library**
- **Canonical Text Services (CTS)**
- **Implementation options**
- **Evaluation and comparison**

Automatische Sprachverarbeitung

The challenge for Digital Humanities

- **Digitization enables access to a vast number of documents (books, manuscripts, papyri, …)**
- **In order to record and analyse content, we need to make texts accessible by machines**
- **„Every preserved word (or fragment) in every edition, manuscript, inscription, and papyrus … is now an object of interest that must possess a unique identifier.“ (Crane et. al. 2012)**
- **Shift from documents to words/phrases/ sentences**

Automatische Sprachverarbeitung

Technological Requirements

- **Access to all levels of documents (metadata, titel, chapters, paragraphs, verses, sentences, words) for**
  - ***search***
  - ***analysis***
  - ***annotation or editing***
- **Open architecture with the goal of a *research infrastructure***

Automatische Sprachverarbeitung

Project „The billion words library“

- Use *available standards* to address all levels of text and to make use of available research infrastsructures (such as CLARIN)
- Partners: University Library Leipzig, ASV, BSV
- Funding by ESF: Group of ten so-called Young Researchers
- Duration: July 2013 – December 2014

Automatische Sprachverarbeitung

Technological Solution: CTS

- **A standard developed in the homermultitext project (www.homermultitext.org), Smith et.al.2009**
- **CTS URNs to identify and retrieve digital representations of texts**
- **URNs serve to associate objects with each other**
- **CTS consists of two parts:**
  - **a URN scheme - can be used to identify texts, passages and abstractions of both**
  - **protocol to find valid URNs and resolve them to text passages**

Automatische Sprachverarbeitung

URN Scheme

CTS URN has the form:

\url{urn:cts:CTSNAMESPACE:WORK:PASSAGE}

- **WORK identifies an instance of a text or abstracts over multiple versions**
- **PASSAGE identifies a section within a text**

Example

urn:cts:demo:shakespeare.sonnets:35.1-35.4

refers to

Line 1 to 4 of Shakespeares Sonnet 35

Automatische Sprachverarbeitung

URN Scheme – More examples

urn:cts:demo:shakespeare.sonnets:

Shakespeares Sonnets

urn:cts:demo:shakespeare.sonnets.de:

german translation

urn:cts:demo:shakespeare.sonnets:35.1

line 1 in sonnet 35

urn:cts:demo:shakespeare.sonnets:35.1-35.5

line 1 to 5 in sonnet 35

urn:cts:demo:shakespeare.sonnets:35.1@grieved-35.5@faults[1]

line 1 word „grieved“ to line 5 first occurence of „faults“

Automatische Sprachverarbeitung

URN Scheme – Use cases

translate passages

urn:cts:demo:shakespeare.sonnets.en:35.1-35.5
urn:cts:demo:shakespeare.sonnets.de:35.1-35.5

show variations of passages

urn:cts:demo:shakespeare.sonnets.en.translit:35.1-35.5
urn:cts:demo:shakespeare.sonnets.en.norm:35.1-35.5

share textpassages with unique persistent Uniform Resource Names (URN)

no worry about correct edition/translation
no need for a physical copy

standardized identifiers for further research

Automatische Sprachverarbeitung

URNs contain 2 types of information

Hierarchical/Structural

encoded in URN
(inner & outer) structure of documents

Flat

located/specified by URN
text
meta information (language)

Suitability with respect to Billion Words Library

Persistency

citable webbased text repository

Scalability

„billion“ words
citable webbased text repository

Usability

easy to set up
minimal dependencies

Automatische Sprachverarbeitung

State of development CTS 5.0

Initial release Feb. 2014, update coming

- **URN** \url{http://www.homermultitext.org/hmt-docs/specifications/ctsurn/}
- **Protocol** \url{http://www.homermultitext.org/hmt-docs/specifications/cts/}
- **no implementation is fully compliant yet with CTS Specifications 5.0**
- **official validator for CTS5.0 to be released soon**
- **Implementations based on rdf, XML, and SQL**

(...)

(...)

Automatische Sprachverarbeitung

Implementation based on SQL (Tiepmar)

- **URNs stored in MySQLs B-Tree → fast search time**
- **LIKE-operator covers most tasks of hierarchical retrieval**

  SELECT .. WHERE urn LIKE urn:cts:demo:shakespeare.sonnets:%
  SELECT .. WHERE urn LIKE urn:cts:demo:shakespeare.sonnets:35.%

- **incremental integer ensures that textchunks are in document-order**
- **passage retrieval reduced to 3 requests for every possible passage**

  get leftest suitable URN
  get rightest suitable URN
  get texts between both

Automatische Sprachverarbeitung

(...)

- (...)

Automatische Sprachverarbeitung

Test of SQL-Implementation with Billion Words

Testdata

1 CTS with 100 000 editions
1 281 272 600 words (min. 3/edition, max. 69118/edition)

Tests

1) List all editions (1 run)
2) Build passage for full edition
   (= biggest passages that can be build)

Results (Milliseconds)

	min	avg	max	sum
1)	1562	1562	1562	1562
2)	24	78	1401	7'890'835
URNs/edition	5	832	3422	83'204'100

Automatische Sprachverarbeitung

Questions, please

Thank you for listening.

Questions?