Sustainability of Text-Technological Resources

Maik Stührenberg, Michael Beißwenger, Kai-Uwe Kühnberger, Harald Lüngen, Alexander Mehler, Dieter Metzing, Uwe Mönnich

Research Group “Text-Technological Modelling of Information”
Overview

- The Research Group *Text-Technological Modelling of Information*
- Aspects of Sustainability
- Sustainability in the Research Group
- Conclusion
The Research Group **Text-Technological Modelling of Information**

The setting:

- Distributed research group funded by the German Research Foundation (DFG)
- Five projects:
  - A2 (Sekimo)
  - A4 (Indogram)
  - B1 (HyTex)
  - C1 (SemDok)
  - C2 (Ontologies)

- Further information: [http://www.text-technology.de](http://www.text-technology.de)
The Sekimo project (A2)

Research topic:
• Integration of heterogeneous linguistic resources
  • Annotations
  • Document grammars
  • Parsers
  • Lexicons
  • Ontologies
  • ...

Domain:
• Anaphoric relations, specifically the resolution of bridging phenomena between definite descriptions
The Sekimo project (A2)

Project methods:

• Implementation of two architectures capable of dealing with multiple annotated texts
  • Prolog fact base (cf. Witt 2004)
  • XML based Sekimo Generic Format (SGF, cf. Stührenberg & Goecke 2008)
• Both architectures can be used for examining relationships between modelling units derived from different annotation layers without the need for an inline markup unification
• Annotation of a corpus of scientific and newspaper articles for training a machine learning anaphora resolution system
The Sekimo project (A2)
The Sekimo project (A2)
The Sekimo project (A2)
The Sekimo project (A2)
The Sekimo project (A2)
The Sekimo project (A2)
The Sekimo project (A2)

Text Technological Modelling of Information

The Sekimo project (A2)
The Sekimo project (A2)

SGF native XML DBMS

XPath/XQuery

Text

XML document

Parser
Discourse Entity Detection
Unification
Semantic Relations
The *Indogram* project (A4)

Research topic:

- Automatic induction of document grammars as models of web genres
The *Indogram* project (A4)

Research topic:

- Automatic induction of document grammars as models of web genres

**Text Documents**

Content: SVM-Induced Semantic Space

Structure: Quantitative Structure Analysis
### The Indogram project (A4)

#### Research topic:
- Automatic induction of document grammars as models of web genres

<table>
<thead>
<tr>
<th>Text Documents</th>
<th>Document Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content:</strong> SVM-Induced Semantic Space</td>
<td><strong>Content:</strong> Vector Space Model</td>
</tr>
<tr>
<td><strong>Structure:</strong> Quantitative Structure Analysis</td>
<td><strong>Structure:</strong> Quantitative Network Analysis</td>
</tr>
</tbody>
</table>
The *Indogram* project (A4)

**Research topic:**
- Automatic induction of document grammars as models of web genres

<table>
<thead>
<tr>
<th><strong>Text Documents</strong></th>
<th>Content: SVM-Induced Semantic Space</th>
<th>Structure: Quantitative Structure Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>![Text Documents Diagram]</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Document Networks</strong></th>
<th>Content: Vector Space Model</th>
<th>Structure: Quantitative Network Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>![Document Networks Diagram]</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Web Documents</strong></th>
<th>Content: Social Ontology (Wikipedia)</th>
<th>Structure: SVM-based Markov Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>![Web Documents Diagram]</td>
<td></td>
</tr>
</tbody>
</table>
Project method:

- Developing of an algorithm for learning the internal structure of web documents as instances of web genres
  - Not only tagging genre labels to web sites or pages but learning genre-related web document structure

Problem:

- There is no typical page-based web genre structure → classical approaches to structure learning cannot be applied
The *HyTex* project (B1)

Research topic:

- Development & evaluation of innovative strategies for automated text-to-hypertext conversion (segmentation and linking)

Project aims:

- Generating hypertext views which provide the selective reader instant access to all textual units necessary for the proper understanding of the current hypertext node (cf. e.g. Lenz & Storrer, 2006)
The HyTex project (B1)

Project methods:

• Coherence-based approach dealing on different markup layers:
  • Document structure
  • Terms and definitions layer
  • Thematic structure layer
  • Cohesion layer

• Development of TermNet, a WordNet-style semantic net describing technical terminology

• Topic-based linking strategies using a GermaNet-based lexical chaining
The *SemDok* project (C1)

Research topic:

- Development of a text parser (discourse parser) for a complex text type
  - Framework: Rhetorical Structure Theory (RST)
  - Corpus: scientific journal articles

Project Methods:

- Use text-technological (XML-based) formalisms, standards, methods
The *SemDok* project (C1)

Discourse representation in the *SemDok* project:

- No semantic representations of…
  - Sentences
  - Text segments
- Primary sources: morphology, syntax, lexicon
  - Punctuation
  - Lexical discourse markers
  - Morphological and syntactic analyses
- Secondary sources: document-/text structure
  - Logical document structure
  - Thematic structure (lexical cohesion, anaphoric structure)
  - Genre-specific text type structure
The *SemDok* project (C1)

Corpus under investigation:

- 47 scientific journal articles, German, from Lingustik Online (~360 000 word forms)
- + 1 web article on hypertext + 1 newspaper article (Die Zeit)
- The development corpus and its annotations will be made available as soon as legal issues are clarified with the publisher of the research articles

<table>
<thead>
<tr>
<th>XML annotation layers</th>
<th>annotated articles (12/2007)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logical document structure (DOC)</td>
<td>47+2</td>
</tr>
<tr>
<td>Genre-specific text type structure (TTS)</td>
<td>47</td>
</tr>
<tr>
<td>Morphological and syntactic structure (CNX; Machinese Syntax Tagger)</td>
<td>47+2</td>
</tr>
<tr>
<td>Discourse markers (DMS)</td>
<td>47</td>
</tr>
<tr>
<td>Rhetorical structure (RST, SemDok-II)</td>
<td>5+2</td>
</tr>
<tr>
<td>Discourse segments (SEG)</td>
<td>5+2</td>
</tr>
<tr>
<td>Anaphoric Structure (CHS) (Project Sekimo)</td>
<td>3+2</td>
</tr>
<tr>
<td>Lexical chains (LC) (Projects HyTex, IndoGram)</td>
<td>1+2</td>
</tr>
</tbody>
</table>
The Ontologies (C2) project

Research topic:

- How can information extraction from structured data be implemented to build an ontology?

Project methods:

- A weak logical formalism is used (description logic) as representation of ontological knowledge
- Certain types of occurring inconsistencies can be resolved by adaptation algorithms (integration engine)

→ The result is an adapted ontology (integrated ontology)
Quite a heterogeneous group of projects, aims and methods...
Sustainability

Aspects of sustainability (cf. Simons 2007):

- Long-term availability of resources
  - Sustainability of primary and secondary data
  - Sustainability of category systems
  - Availability of methods and tools
- Process properties in a long-term perspective
- Community of experts and non-experts
Sustainability in the Research Group

Aspects of sustainability (cf. Simons 2007):

• Long-term availability of resources
  • Sustainability of primary and secondary data
  • Sustainability of category systems
  • Availability of methods and tools
• Process properties in a long-term perspective
• Procedural aspects of sustainability
• Community of experts and non-experts
Sustainability in the Research Group

Aspects of sustainability (cf. Simons 2007):

• Long-term availability of resources
  • Sustainability of primary and secondary data
  • Sustainability of category systems
  • Availability of methods and tools
• Process properties in a long-term perspective
• Procedural aspects of sustainability
• Community of experts and non-experts
Sustainability in the Research Group – Primary and secondary data

Primary data

• Several projects share the same corpus (as a whole or in parts)
  • Corpus of scientific articles with logical document layer developed by B1 & C1 → minimisation of corpus work
  • Corpus at least partially available online, Options guaranteeing an open access (Open Access or Creative Commons) are under discussion
  • For automatic annotation of syntactic and dependency features the same commercial parser is used (Machinese Syntax by Connexor Oy) → same base layer for further annotation efforts
  • All corpora are encoded in XML, either compatible with existing standards (e.g. DocBook) or specified by document grammars (DTD or XSD) that are available online
  • Projects A4 & C1 share the same architecture → combination of annotation efforts possible
Primary data – the generic format SGF (cf. Stührenberg & Goecke 2008)

• Based on Annotation Graphs, but uses the XML-inherent tree structures whenever possible to reduce processing costs

• Standoff annotation allowing for several annotation layers over the same primary data in a single file

• Import scripts transforming arbitrary inline annotation into SGF layers and combining them afterwards into a single SGF file while most parts of the original annotation is conserved → no loss of information

• Uses standard XPath/XSLT/XQuery for processing/querying

• Use on a per-file base, native XML database or relational database (MySQL API available)

• Uses XML Schema and additional Schematron constraints, relies heavily on XML namespaces

• Uses checksum mechanisms to maintain identify of primary data through the annotation process
Primary data – the generic format SGF (cf. Stührenberg & Goecke 2008)
Primary data – the generic format SGF (cf. Stührenberg & Goecke 2008)
Secondary data

- Metadata has been applied to all corpus data
  - OLAC (A2 exclusively)
  - OLAC with additional elements (C1)
Aspects of sustainability:

- Long-term availability of resources
  - Sustainability of primary and secondary data
  - Sustainability of category systems
  - Availability of methods and tools
- Procedural aspects of sustainability
- Process properties in a long-term perspective
- Community of experts and non-experts
Sustainability in the Research Group

Aspects of sustainability:

- Long-term availability of resources
  - Sustainability of primary and secondary data
  - Sustainability of category systems
  - Availability of methods and tools
- Process properties in a long-term perspective
- Procedural aspects of sustainability
- Community of experts and non-experts
Sustainability of category systems

- Category systems (like Ontologies) can be used as mediator between different sets of annotation (cf. Schmidt et al. 2006), however, one has to assure that the category system is sustainable as well.
- Usage of standardized formal ontologies specifying the categories (OWL).
- Reference ontologies as instruments to ensure the sustainability of linguistic tag sets (e.g. GOLD).
- (Re-)construction of ontological resources in a standardized ontology formalism.
Sustainability of category systems

Cooperation between Projects B1 & C1 and the GermaNet group in Tübingen

• Aims: Making (Princeton-style) general-language word nets and domain-specific word nets interoperable – and thus sustainable –, by
  • Constructing different representations of word nets using the Web Ontology Language (OWL)
  • Experimenting with various modelling options (OWL DL, OWL FULL)
  • Developing an approach for an OWL-based linking of general language and terminological word nets
Cooperation between Projects B1 & C1 and the GermaNet group in Tübingen

- Current results/resources:
  - Subsets of GermaNet in OWL
  - TermNet (terminological lexical resource on the basis of a word net model that has been extended for terminologies; cf. Beißwenger, 2008; Selzam, 2008) in OWL (two alternative representations)
  - GermaTermNet in OWL (TermNet & subset of GermaNet connected in OWL by means of a set of “plug-in relations”; cf. Lüngen and Storrer, 2007; Kunze et al., 2007; Lüngen et al., 2008) (two alternative representations)
- All resources are freely available online
Sustainability in the Research Group

Aspects of sustainability:

• Long-term availability of resources
  • Sustainability of primary and secondary data
  • Sustainability of category systems
  • Availability of methods and tools

• Procedural aspects of sustainability

• process properties in a long-term perspective

• Community of experts and non-experts
Sustainability in the Research Group

Aspects of sustainability:

• Long-term availability of resources
  • Sustainability of primary and secondary data
  • Sustainability of category systems
  • Availability of methods and tools

• Process properties in a long-term perspective

• Procedural aspects of sustainability

• Community of experts and non-experts
Availability of methods and tools

• Documentation of markup specifications and annotation guidelines is freely available online
  • Thematic structure
  • Coreference phenomena
  • Term definitions
  • Anaphoric relations

• Several tools are freely available online
  • Scientific Workplace/Desktop
  • Ariadne System
  • Serengeti Web annotation tool
  • Lexical Chainer GLexi
  • GermaNet Explorer
  • GermaNet Pathfinder
Availability of methods and tools – Scientific Workplace/Desktop
Main application for handling text (txt, html, pdf) documents, including:

- Corpus management
- Linguistic tagger for logical document structure
- Lemmatization
- NE recognition
- Lexical chaining
Sozial ist schlau
VON ULRICH BAHNSEN


Primaten mit Plan - Wissen Affen, was Zukunft ist?
Ein Affe will Rechte - Darf ein Schimpanse vor Gericht ziehen?
Availability of methods and tools – Scientific Workplace/Desktop

From Content Analysis to Semantic Search
Availability of methods and tools – Scientific Workplace/Desktop

From Content Analysis to Semantic Search

Lexeme Cloud

XHTML

PREMIATENFORSCHUNG
Sozial ist schlau
Warum können Menschen besser denken als Affen? Eine neue wissenschaftliche Studie belegt: weil sie besonders soziale Wesen sind.

Primaten mit Plan - Wissen offen, was Zukunft ist
Ein Affe will Rechte - Drei an Schwierigkeiten vor Gericht stehen
From Content Analysis to Semantic Search

Lexeme Cloud → Terminology Explorer

Availability of methods and tools – Scientific Workplace/Desktop

Kategorie: Lernen

In dieser Kategorie wird alles aufgelistet, was das Lernen betrifft.

Unterkategorien

Es werden 7 Unterkategorien angezeigt.

B
- [*] Berufliches Lernen
L
- [*] Lernpsychologie
S
- [*] Sprachenlernen

E
- [*] E-Learning
M
- [*] Maschinelles Lernen
- [*] Mnemotechnik

[*] Studiertechnik
Availability of methods and tools – Scientific Workplace/Desktop

From Content Analysis to Semantic Search

Lexeme Cloud → Terminology Explorer → Concept Cloud

Lexeme Cloud

Terminology Explorer

Concept Cloud

CategoryTagCloud (U10)

Allgemeine Psychologie Pädagogische Psychologie Gedächtnis Neurobiologie Lernen Kognitive Leistungssteigerung

LemmaArticleTagCloud (U11)

Lernen (100.0%) Denken (93.1%) Lernpsychologie (92.6%)
Availability of methods and tools – Scientific Workplace/Desktop

From Content Analysis to Semantic Search

- Lexeme Cloud
- Terminology Explorer
- Concept Cloud
- Semantic Search

[XHTML]

Denken

Search:
Denken
Title: Grenze des Denkens ; 257
Info:
Content: ...Autoren A - Z | Startseite | Info | Shop | Genres | Neu | Impressum | Mitmachen | Druckversion Autoren Grenze des Denkens. Rein zu denken, vermeinst du? Wie sehr du die Bilder auch sichtest, Bilder bleiben es doch, wenn auch zu Worten verdünnt, Und...
Author: Friedrich Hebbel
Link: http://gutenberg.spiegel.de/hebbel/gedichte/denken.htm
Availability of methods and tools – Scientific Workplace/Desktop

From Content Analysis to Semantic Search

Implication:

Automatic generation of semantic searches
Availability of methods and tools – Ariadne System
Availability of methods and tools – Ariadne System

- Built-up, maintenance, annotation and statistical analysis of text, web and multimodal data as well as document networks in a single system
- User management system for organizing larger annotation projects
- Inclusion of the HyGraphDB as a linguistic database based on graph theory
  HyGraphDB is based on the Berkeley DB and utilizes the Graph eXchange Language (GXL) to represent linguistic data structure
Das WWW\textsuperscript{34} ist nicht die einzige Möglichkeit\textsuperscript{35}, einen Hypertext\textsuperscript{36} zu realisieren, es gib zahlreiche Hypertextsysteme\textsuperscript{38} (z.B. Hypercard\textsuperscript{39}, Toolbook\textsuperscript{40}), mit denen sich lokale, nicht übermehrere Computer\textsuperscript{43} vernetzte Hypertexte\textsuperscript{44} generieren lassen (z.B. für CD-ROM-Anwendungen\textsuperscript{45}). Auf die Unterschiede\textsuperscript{46} zwischen den einzelnen Hypertextsystemen\textsuperscript{47} soll hier nicht eingegangen werden.

Nur noch ein Wort\textsuperscript{48} zu den Begriffen "Hypermedia" und "Multimedia"\textsuperscript{49}. "Hypertext"\textsuperscript{52} umfaßt nur ein "Medium"\textsuperscript{53}, nämlich das Medium "Text"\textsuperscript{54}. Sobald weitere Medien\textsuperscript{56} hinzu kommen (Graphik\textsuperscript{57}, Bilder\textsuperscript{58}, Ton\textsuperscript{59}, Video\textsuperscript{60}), wird daraus eine Hypermedia-Anwendung\textsuperscript{51}. Eine Diskussion\textsuperscript{53} des Begriffs "Multimedia"\textsuperscript{52} finden Sie\textsuperscript{55} hier. "Multimedia"\textsuperscript{55} bedeutet - wenn man sich\textsuperscript{58} mit einer einfachen Erklärung\textsuperscript{59} zufriedengeben will - "Einsatz\textsuperscript{70} mehrerer Medien\textsuperscript{71}" ohne Rücksicht\textsuperscript{72} darauf, ob "Informationseinheiten\textsuperscript{73}" vorliegen, die\textsuperscript{74} mit Hyperlinks\textsuperscript{75} miteinander verknüpft sind. Die Begriffe "Hypermedia" und "Multimedia"\textsuperscript{76}" darf man\textsuperscript{79} also nicht in einen Topf\textsuperscript{83} werfen, obwohl dies\textsuperscript{83} häufig geschieht.

1. Text\textsuperscript{82}, Hypertext\textsuperscript{83}, Wissensbank\textsuperscript{84}
Availability of methods and tools – Serengeti Web Annotation Tool

- Login/logout
- Choose group/document (and annotation if you have the right to)
- Save/export/print annotation
Availability of methods and tools – Serengeti Web Annotation Tool


1. Text, Hypertext, Wissensbank

Annotate the text:
- choose the discourse entities that take part in a relation
Availability of methods and tools – Serengeti Web Annotation Tool

Annotate the text:
• choose the discourse entities that take part in a relation
• choose the relation type and subtype
Availability of methods and tools – Serengeti Web Annotation Tool

Annotate the text:
• choose the discourse entities that take part in a relation
• choose the relation type and subtype
• comment if you're unsure
Availability of methods and tools – Serengeti Web Annotation Tool

- Screenshots taken from the current stable version
- In July the next version of Serengeti will be available, with the following changes:
  - Input text either raw or pre-annotated
  - Editing of markables (adding, deleting, changing a markable's attributes)
  - Import & Export from and to the Sekimo Generic Format (SGF)
- While the Scientific Workplace/Desktop are freely available for online use, Serengeti will go Open Source after the end of the project
Sustainability in the Research Group

Aspects of sustainability:

• Long-term availability of resources
  • Sustainability of primary and secondary data
  • Sustainability of category systems
  • Availability of methods and tools

• Process properties in a long-term perspective

• Procedural aspects of sustainability

• Community of experts and non-experts
Sustainability in the Research Group

Aspects of sustainability:

• Long-term availability of resources
  • Sustainability of primary and secondary data
  • Sustainability of category systems
  • Availability of methods and tools
• Process properties in a long-term perspective
• Procedural aspects of sustainability
• Community of experts and non-experts
Procedural aspects of sustainability

- Sustainability cannot be assured only by a static saving of data
- Therefore, dynamic processes should be used to allow for a wider scope of sustainability
  - Learning and induction of ontological systems (e.g. as seen in the Scientific Workplace/Desktop system)
  - Consistency preservation of ontological resources
  - Formal properties of data structures (e.g. algorithmically transformation of Annotation Graphs into multi-rooted trees, cf. Michaelis and Mönnich 2007)
Aspects of sustainability:

- Long-term availability of resources
  - Sustainability of primary and secondary data
  - Sustainability of category systems
  - Availability of methods and tools
- Process properties in a long-term perspective
- Procedural aspects of sustainability
- Community of experts and non-experts
Sustainability in the Research Group

Aspects of sustainability:

• Long-term availability of resources
  • Sustainability of primary and secondary data
  • Sustainability of category systems
  • Availability of methods and tools
• Process properties in a long-term perspective
• Procedural aspects of sustainability
• Community of experts and non-experts
Community of experts and non-experts

Existence of a community and a complex network of valuable co-operation as central aspect of sustainability

• As already seen, co-operation between projects in the research group is widely accepted

• Co-operation between the A2 project and the international projects AnaWiki and AnaphoricBank
  → contributing to its corpora, representation format and architecture (Serengeti)

• Wiki implemented in the A4 project that supports collaboration with respect to the exchange and maintenance of treebanks

• As already said, most of the tools developed are available online together with a comprehensive documentation, allowing users to contribute
Conclusion

Sustainability has many facets, but...

- ...even in a heterogeneous research group such as ours, minimization of effort is possible by using sustainable data, resources, methods and tools
- ...the publicly available access to most of the documentation and tools can help other projects and research group in the same way and can contribute to a community building process (in principle, cf. “not invented here” problem)

Please try not to reinvent the wheel
Thank you for your attention!

http://www.text-technology.de/
maik.stuehrenberg@uni-bielefeld.de