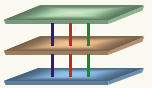
Generating topic chains and topic views: Experiments using GermaNet

Irene Cramer, Marc Finthammer, and Angelika Storrer
Faculty of Cultural Studies,
Dortmund University of Technology, Germany

irene.cramer@uni-dortmund.de

Outline



- Project context
- Concept of topic chains and topic views
- Construction of topic views
- Evaluation

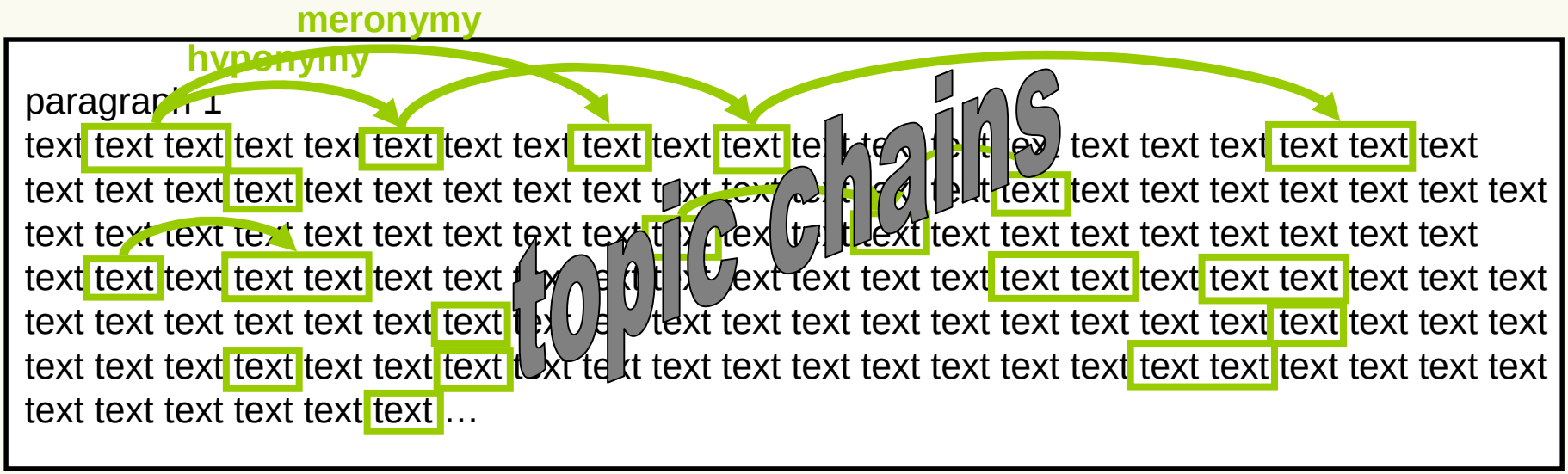
Project context

Text Technological Modelling of Information

- Project *HyTex* on text-grammatical foundations for the (semi-)automated text-to-hypertext conversion
- Research line in this context: topic-based linking strategies using lexical chaining as a resource

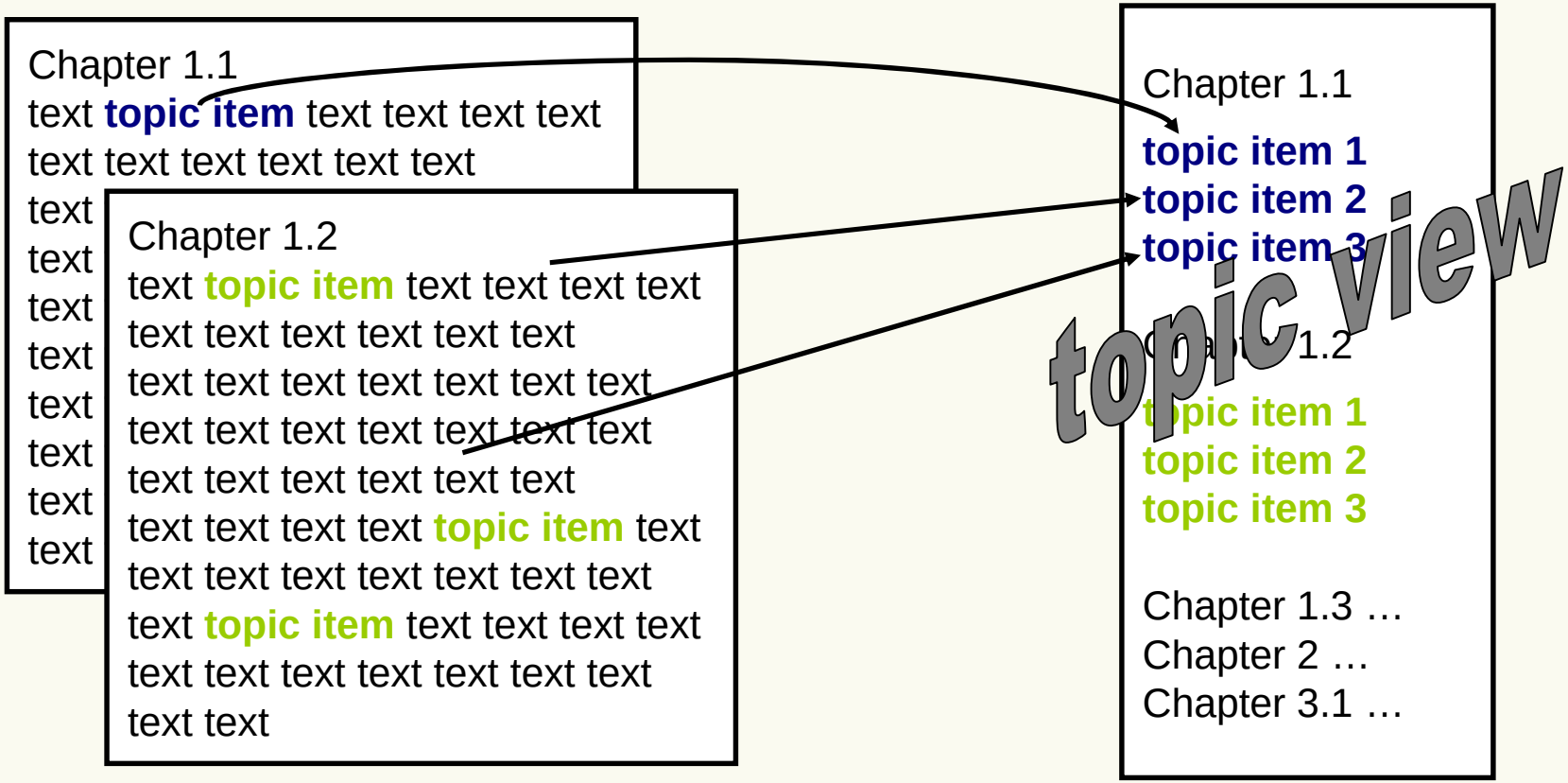
Concept of topic chains

- Partial text representation based on selection of thematically central words (multi word units), so called topic items
- Instrument for visualization of thematic development in text segments (meant as analysis tool for linguists)



Concept of topic views

- Thematic index based on text grammatical information constructed of a selection of topic items
- Intended to support the user's orientation and navigation



Example: Kinderarmut

Text Technological Modelling of Information

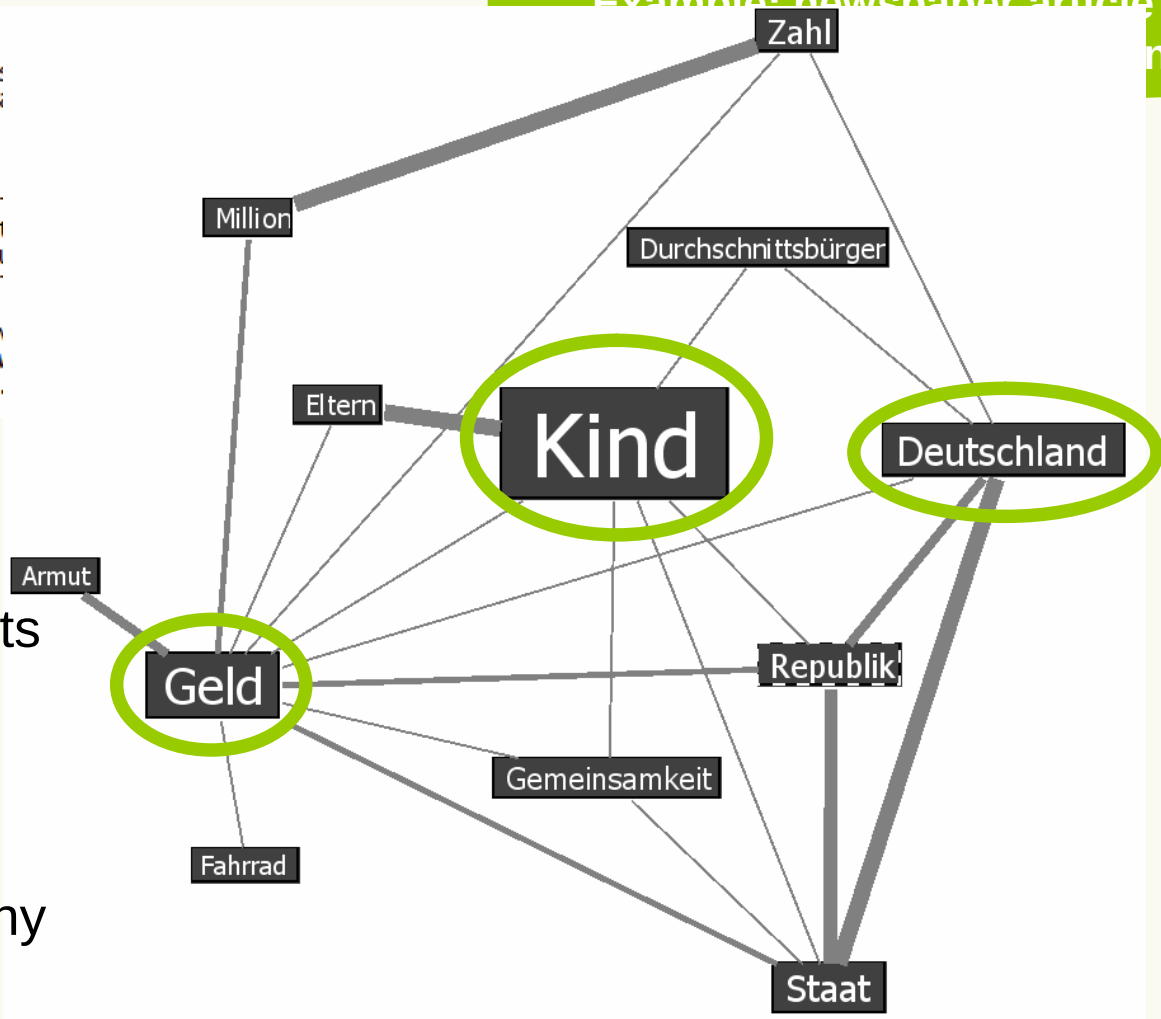
Example: newspaper article

Arm, ärmer, Kind

Die Zahl der Kinder, die Not leiden, i
schneller gestiegen als in fast allen

Von Philipp Krohn

Maik hat kein eigenes Fahrrad, und Mar
Urlaub. Zwei Kinder, eine Gemeinsamkei
viel Geld im Monat wie ein deutscher D
sie als "arm" - zwei von rund 1,5 Mill
die gemeint sind, wenn von Kinderarmut
sind nicht obdachlos. Sie sehen nicht
benachteiligt - ganz einfach deshalb,
meisten anderen Kinder in der Republik.



Topic items according method in initial experiments

Kind, Engl. child

Geld, Engl. money

Deutschland, Engl. Germany

Construction of topic chains and views

Text Technological Modelling of Information

1-3 best topic items per paragraph

à topic view

output lexical chaining + additional features

↕

net representing with topic items as nodes and semantic relatedness as edges

use all topic items per paragraph for another chaining step

à topic chain

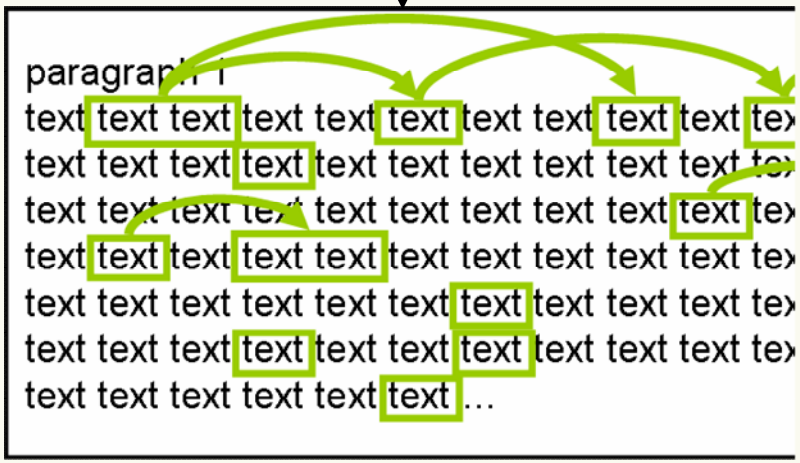
Chapter 1.1

topic item 1
topic item 2
topic item 3

Chapter 1.2

topic item 1
topic item 2
topic item 3

Chapter 1.3 …
Chapter 2 …
Chapter 3.1 …

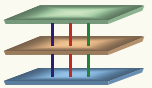


HyTex – Hypertextualisierung auf textgrammatischer Grundlage

http://www.hytex.info/

tu

Basis of our approach: *GLexi*



Our lexical chainer *GLexi*:

- **modular architecture**:
  - § linguistic preprocessing, XML-input
  - § core algorithm with interface for integration of various resources / relatedness measures
  - § output generation – several formats (e.g. XML and visual graph representation)
- **evaluation** wrt. coverage, disambiguation quality, performance of semantic relatedness measures, and application (see Cramer & Finthammer, 2008)

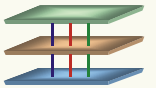
Basis of our approach: GLexi

Text Technological Modelling of Information

Principle parameter settings:	**Parameter setting for construction of topic views / topic chains***
• preprocessing (using TEMIS tools): lemmatization, morphological analysis, POS-Tagging	• all preprocessing steps (using TEMIS tools)
• resources: GermaNet, GermaTermNet (extension of GN with terminology), Google co-occurrence counts	• GermaTermNet
• 11 semantic relatedness measures: - 8 based on GermaNet (or GermaTermNet): Graph-Path, Tree-Path, Wu-Palmer (1994), Leacock-Chodorow (1998), Hirst-StOnge (1998), Resnik (1995), Jiang-Conrath (1997), Lin (1998) - 3 based on Google: Google-Quotient, Google-NDG, Google-PMI	• Lin's measure (based on GermaTermNet) thresholds 0.4 not_related – related and 0.7 related – strongly related

* decision based on experiments reported in Cramer & Finthammer, 2008

Construction of topic views - overview



Intuition – topic item:

- § lexical item central for topic(s) in paragraph

Automatic selection of topic items:

- § select relevant lexical items per paragraph
  à topic item candidates, called TIC
- § build network with TICs as nodes and weighted edges based on GLexi
- § remove edges with low relatedness values (according to our threshold)
- § select topic items

Construction of topic views - overview

Text Technological Modelling of Information

Criteria for topic item selection:

- parameters:
  - § relative frequency,
  - § density in TIC-net,
  - § relation strength in TIC-net
- use linear combination to calculate topic relevance values for each TIC using these parameters
- derive ranking on basis of topic relevance values

Construction of topic views – step-by-step

Text Technological Modelling of Information

paragraph 1
text text text text text text text
text text text text text text
text text text text text text text
text text text text text
text text text text text text
text text text text text text text
text text text text text
text text text text text text text
text text text …

paragraph 2
text text text text text text text
text text text text text
text text text text text text
text text text text text text text
text text text text text text
text text text text text text text
text text …

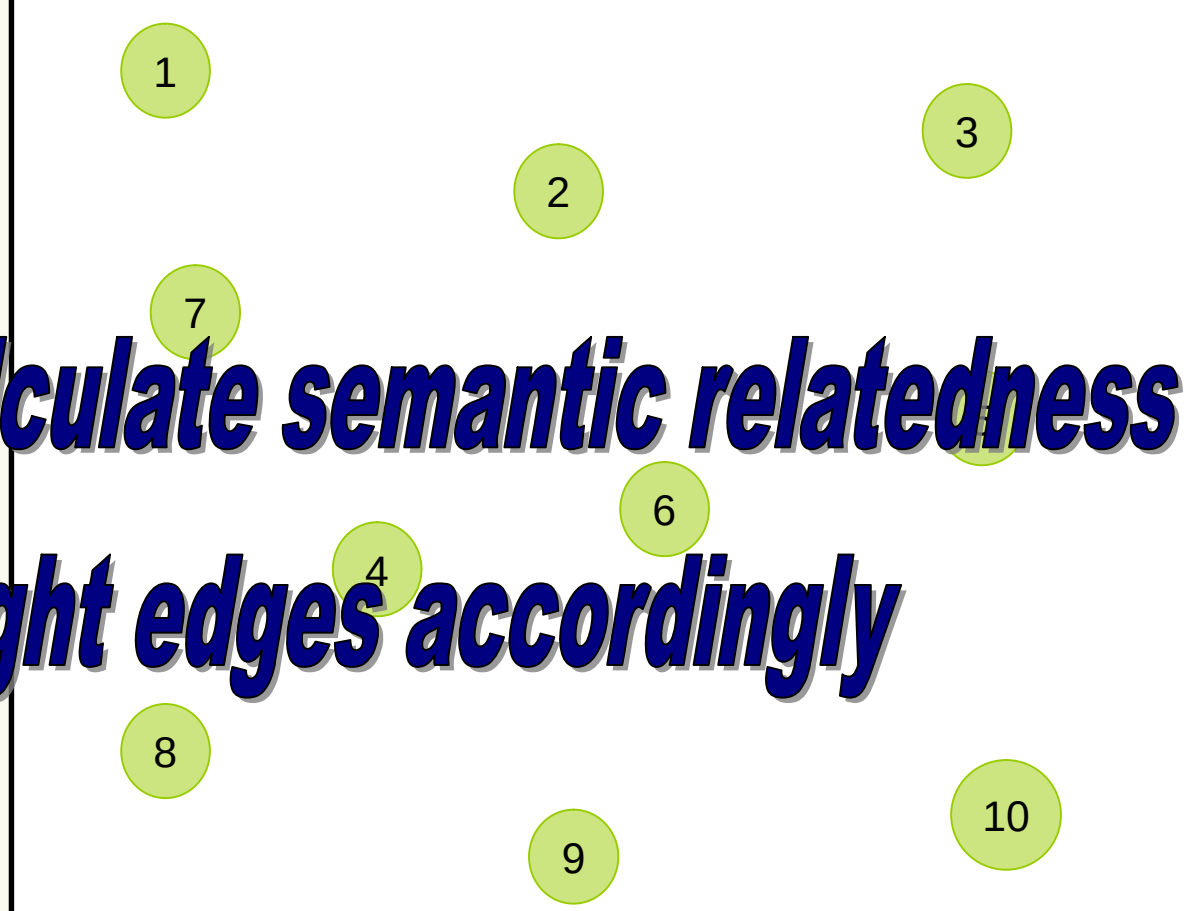
Step 1: find topic item candidates (TIC): thematically central lexical unit

Construction of topic views – step-by-step

Text Technological Modelling of Information

paragraph 1
text text text text text text text
text text **tic** text text
text text text text text text text
text text text text **tic** text text
text text text text text text
text text text **tic** text text
text text text text text text text
text **tic** text text text text
text text text text text text text
text text text **tic** …

paragraph 2
text text text text **tic** text
tic text text text text
text text text text text text
text text text **tic** text
text text **tic** text text
text text text text text text text
text **tic** …



Step 2: calculate semantic relatedness

weight edges accordingly

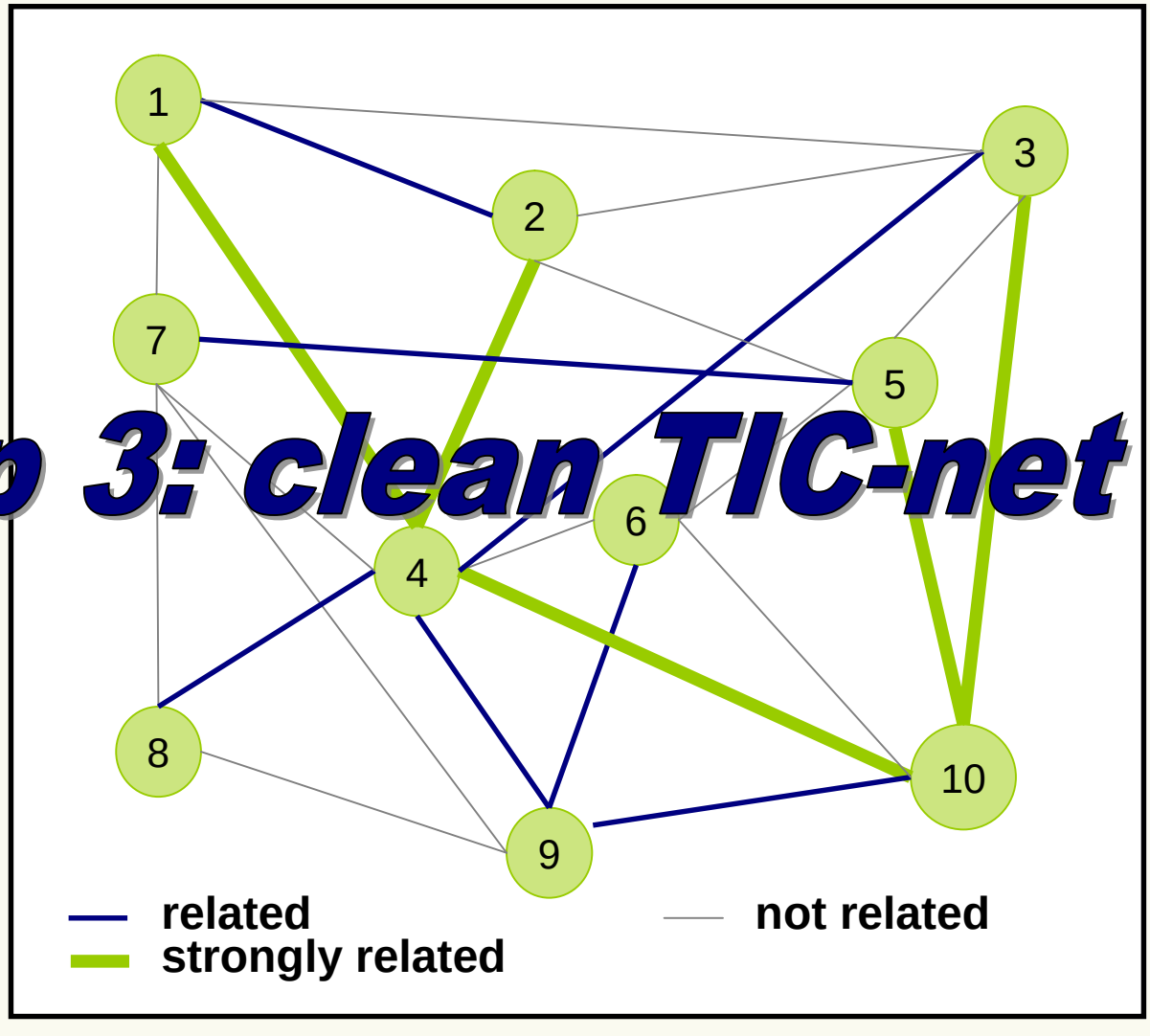
1 … 10 tic (= topic item candidates)

Construction of topic views – step-by-step

Text Technological Modelling of Information

paragraph 1
text text text text text text text
text text **tic** text text
text text text text text text text
text text text text **tic** text text
text text text text text text
text text text **tic** text text
text text text text text text text
text **tic** text text te
text text text text
text text text **tic** …

paragraph 2
text text text text **tic** text
tic text text text text
text text text text text text
text text text **tic** text
text text **tic** text text
text text text text text text text
text **tic** …



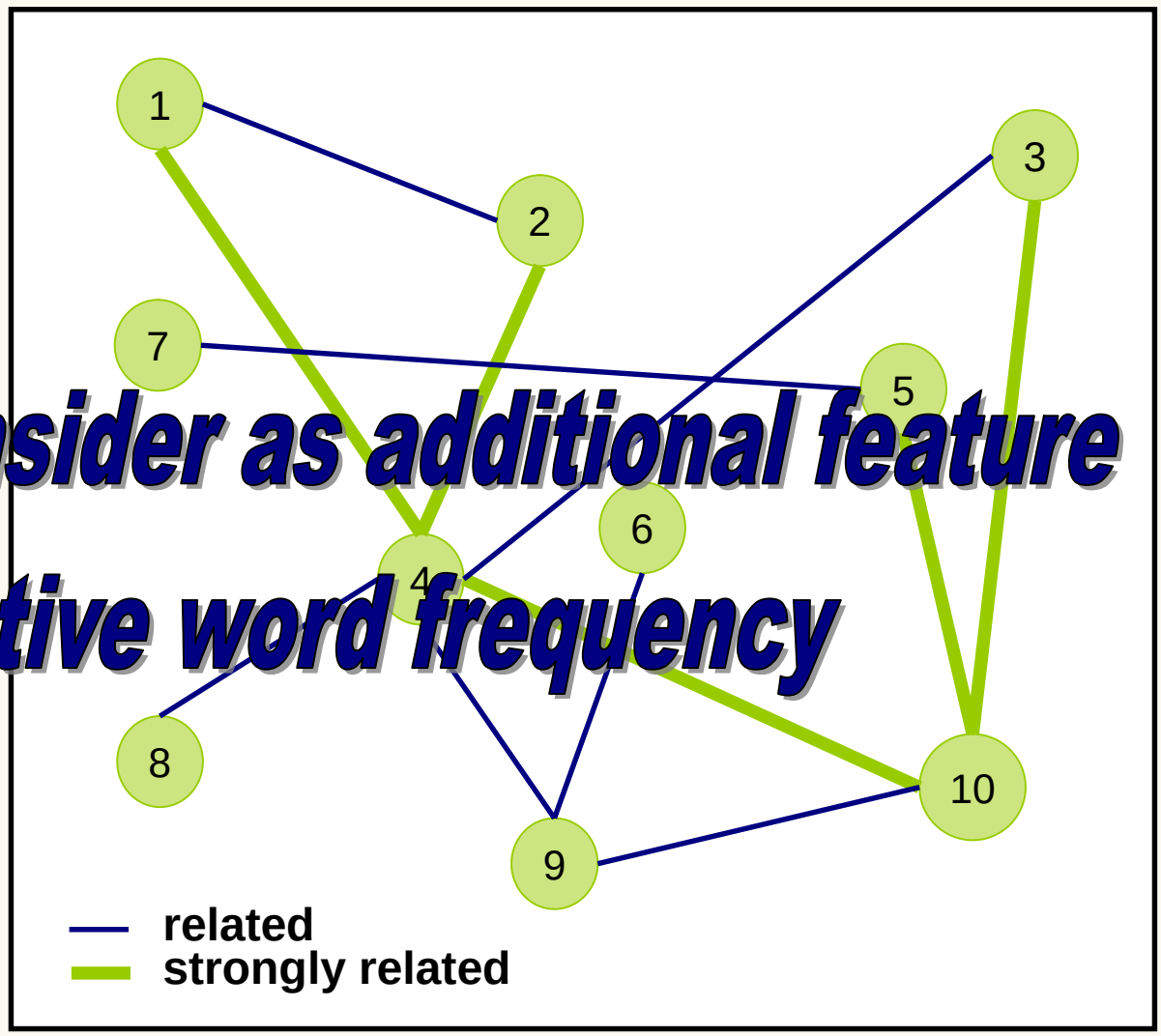
Step 3: clean TIC-net

Construction of topic views – step-by-step

Text Technological Modelling of Information

paragraph 1
text text text text text text text
text text **tic** text text
text text text text text text text
text text text text **tic** text text
text text text text text text
text text text **tic** text text
text text text text text text text
text **tic** text text
text text text text text text text
text text text **tic** …

paragraph 2
text text text text **tic** text
tic text text text text
text text text text text text
text text text **tic** text
text text **tic** text text
text text text text text text text
text **tic** …



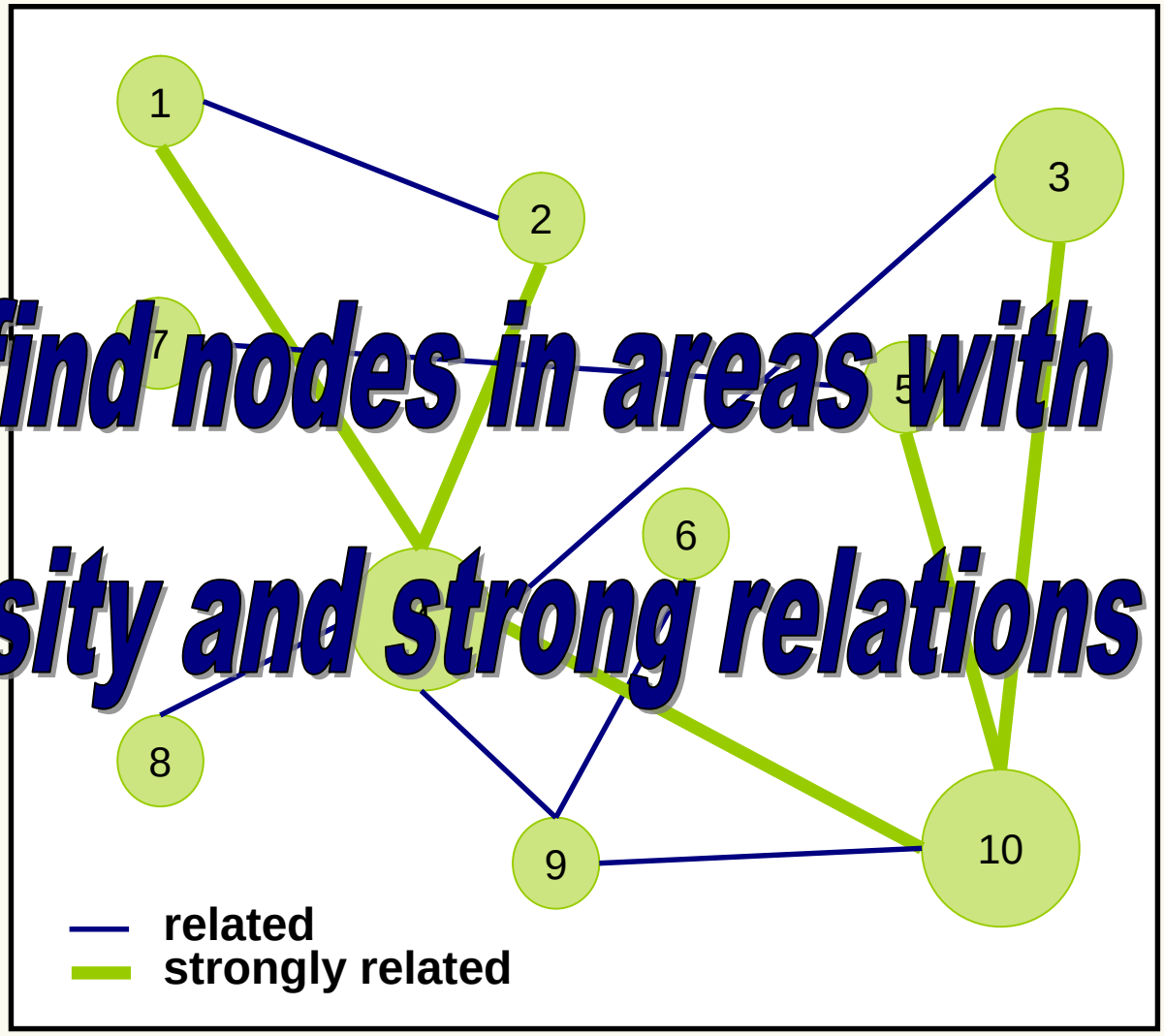
Step 4: consider as additional feature relative word frequency

Construction of topic views – step-by-step

Text Technological Modelling of Information

paragraph 1
text text text text text text text
text text **tic** text text
text text text text text text text
text text text text **tic** text text
text text text text text text
text text text text text text
text text text text text text
text tic text text text
text text text text text text text
text text text **tic** …

paragraph 2
text text text text **tic** text
tic text text text text
text text text text text text
text text text **tic** text
text text **tic** text text
text text text text text text text
text **tic** …



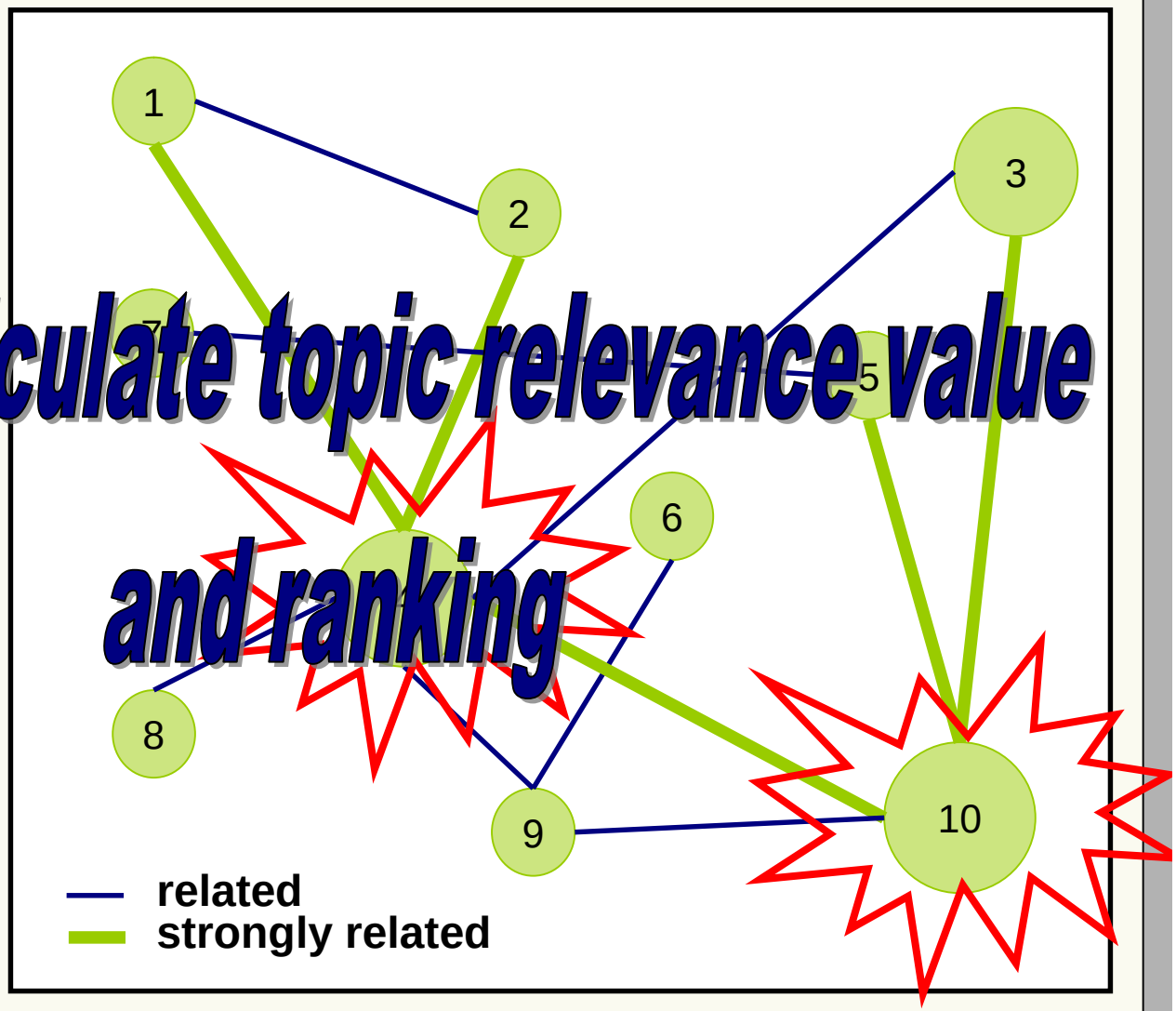
Step 5: find nodes in areas with high density and strong relations

Construction of topic views – step-by-step

Text Technological Modelling of Information

paragraph 1
text text text text text text text
text text **tic** text text
text text text text text text text
text text text text **tic** text text
text text text text text text
text text text text text text
text text text text text text
tic text text text
text text text text text text text
text text text **tic** …

paragraph 2
text text text text **tic** text
tic text text text text
text text text text text text
text text text **tic** text
text text **tic** text text
text text text text text text text
text **tic** …



Step 6: calculate topic relevance value and ranking

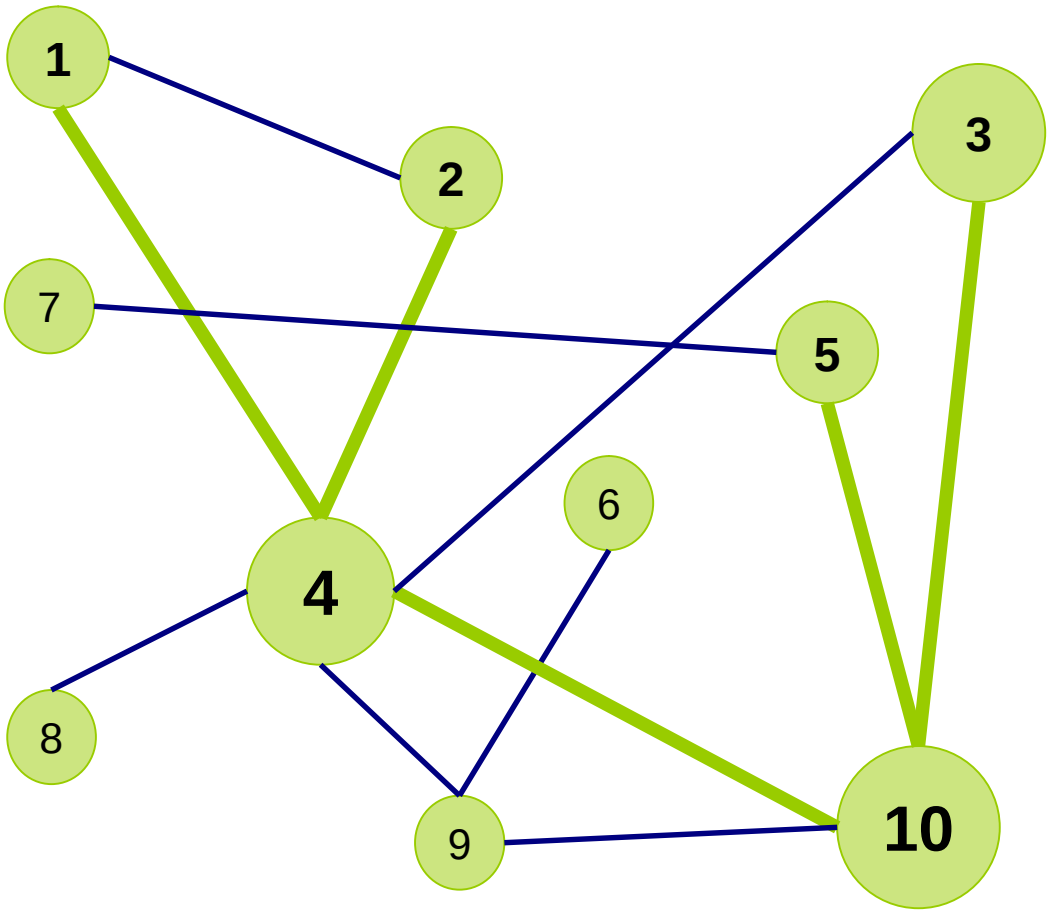
Construction of topic views – step-by-step

Text Technological Modelling of Information

paragraph 1
text text text text text text text
text text **tic** text text
text text text text text text text
text text text text **tic** text text
text text text text text text
text text text **tic** text text
text text text text text text text
text **tic** text text text
text text text text text text text
text text text **tic** …

3 criteria for selection + ranking:

1. relative frequency
2. density
3. relation strength

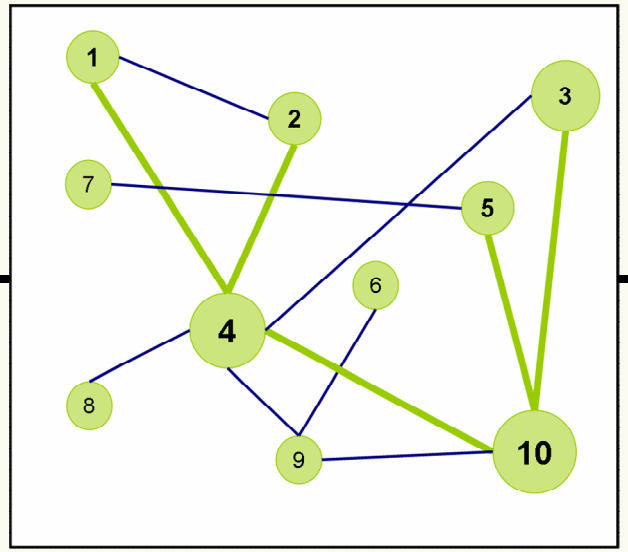


Construction of topic chains and views

Text Technological Modelling of Information

1-3 best topic items per paragraph

à topic view



use all topic items per paragraph for another chaining step

à topic chain

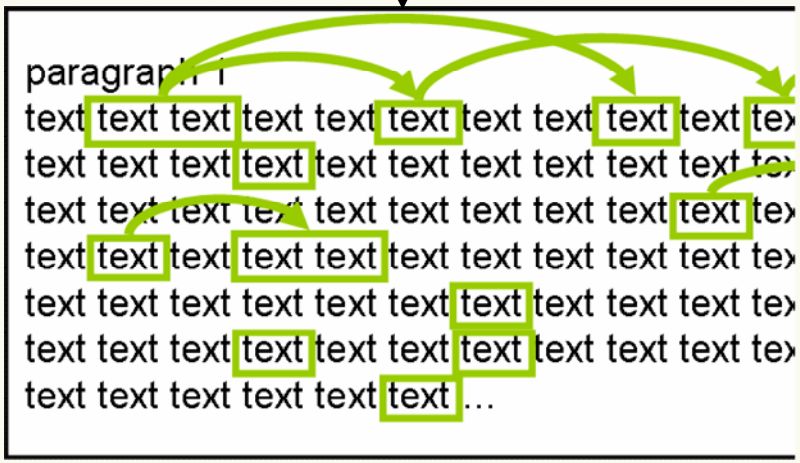
Chapter 1.1

topic item 1
topic item 2
topic item 3

Chapter 1.2

topic item 1
topic item 2
topic item 3

Chapter 1.3 …
Chapter 2 …
Chapter 3.1 …



Evaluation

Text Technological Modelling of Information

- Manual annotation of topic items in part (80 paragraphs) of HyTex core corpus (annotator agreement: approx. 70 %)

  à gold standard for evaluation of automatic extraction

- Automatic extraction of topic items in part (107 paragraphs) of HyTex core corpus

overlap with manual annotation	1	≤ 2/3 and ≥ 1/3	0	≥ 1/3
Doc 1 (39 par.)	23 %	59 %	18 %	82 %
Doc 2 (49 par.)	18 %	53 %	29 %	71 %
Doc 3 (19 par.)	53 %	32 %	16 %	85 %
mean all 3 docs	31 %	48 %	21 %	79 %

Outlook

Text Technological Modelling of Information

Observations:

- if all relevant words in the paragraph are appropriately represented in the lexical semantic resource
  … then performance of automatic topic item extraction is good
- the longer a paragraph, the better the extraction of topic items

Challenges in automatic topic item extraction:

- Named Entities
- technical terms
- multi word preprocessing

Plans:

- integration of topic views as new navigation tool into HyTex demo prototype
- experiments on refinement of manual annotation and automatic extraction, especially, more features in TIC selection such as mark-up and tf/idf-methods for density and strength

Thank you!

More information about our research can be found at our project web-pages:

www.hytex.info