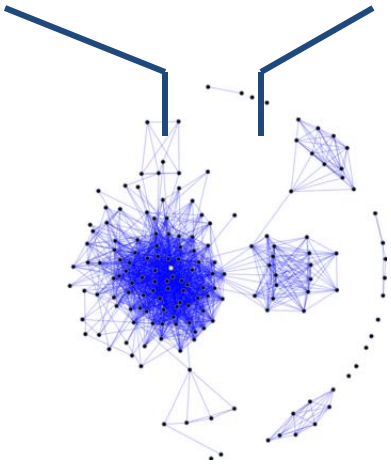
Automated Text Analysis for eHumanities

TECHNISCHE UNIVERSITÄT DARMSTADT

Silvana Hartmann



UKP Lab: Snapshot of January 2011

TECHNISCHE UNIVERSITÄT DARMSTADT

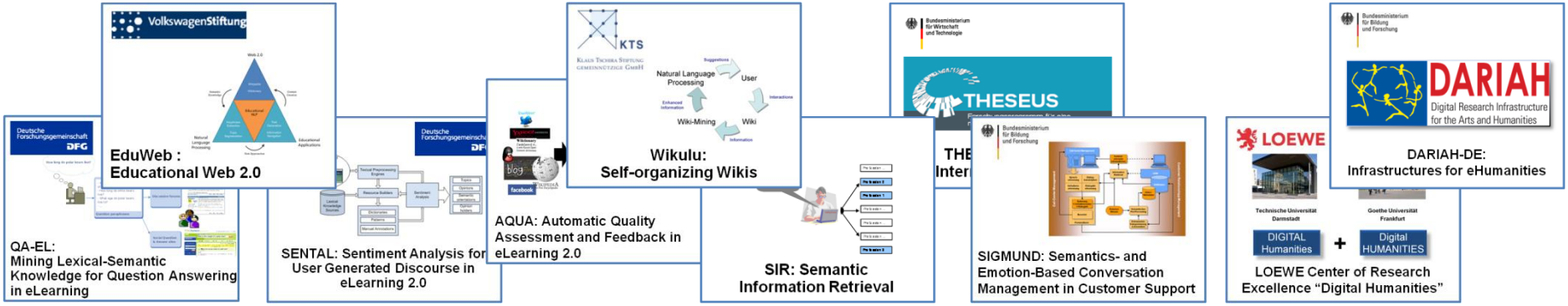
- 1 Junior Prof.
- 3 junior research group leaders
- 2 PostDocs
- 14 PhD students

UKP Lab: Projects

TECHNISCHE UNIVERSITÄT DARMSTADT

UKP Lab

- **Educational Natural Language Processing**
- **Natural Language Processing and Wikis (Dr. Torsten Zesch)**
- **Semantic Information Management (Dr. György Szarvas)**
- **Language Technology for eHumanities (Richard Eckart, Dipl. Inf.)**
- **Statistical Semantics (Jun. Prof. Chris Biemann)**



UKP

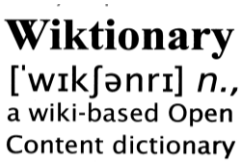
UKP Lab: Software and Data

TECHNISCHE UNIVERSITÄT DARMSTADT

- Darmstadt Knowledge Processing (DKPro) Repository

- Wikipedia API

- Wiktionary API



- OmegaWiki API

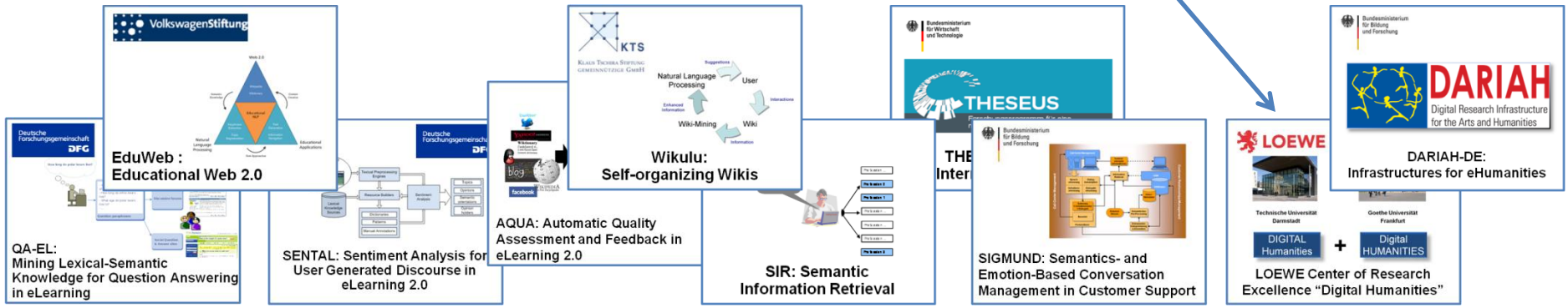
http://www.ukp.tu-darmstadt.de/software/

UKP Lab: Projects

TECHNISCHE UNIVERSITÄT DARMSTADT

UKP Lab

- **Educational Natural Language Processing**
- **Natural Language Processing and Wikis (Dr. Torsten Zesch)**
- **Semantic Information Management (Dr. György Szarvas)**
- **Language Technology for eHumanities (Richard Eckart, Dipl. Inf.)**
- **Statistical Semantics (Junior-professur)**



UKP

UKP Lab: Language Technology for eHumanities

TECHNISCHE UNIVERSITÄT DARMSTADT

Bundesministerium für Bildung und Forschung

DARIAH
Digital Research Infrastructure for the Arts and Humanities

DARIAH-DE:
Infrastructures for eHumanities

LOEWE

Technische Universität Darmstadt

Goethe Universität Frankfurt

DIGITAL Humanities + Digital HUMANITIES

LOEWE Center of Research Excellence "Digital Humanities"

DARIAH-DE

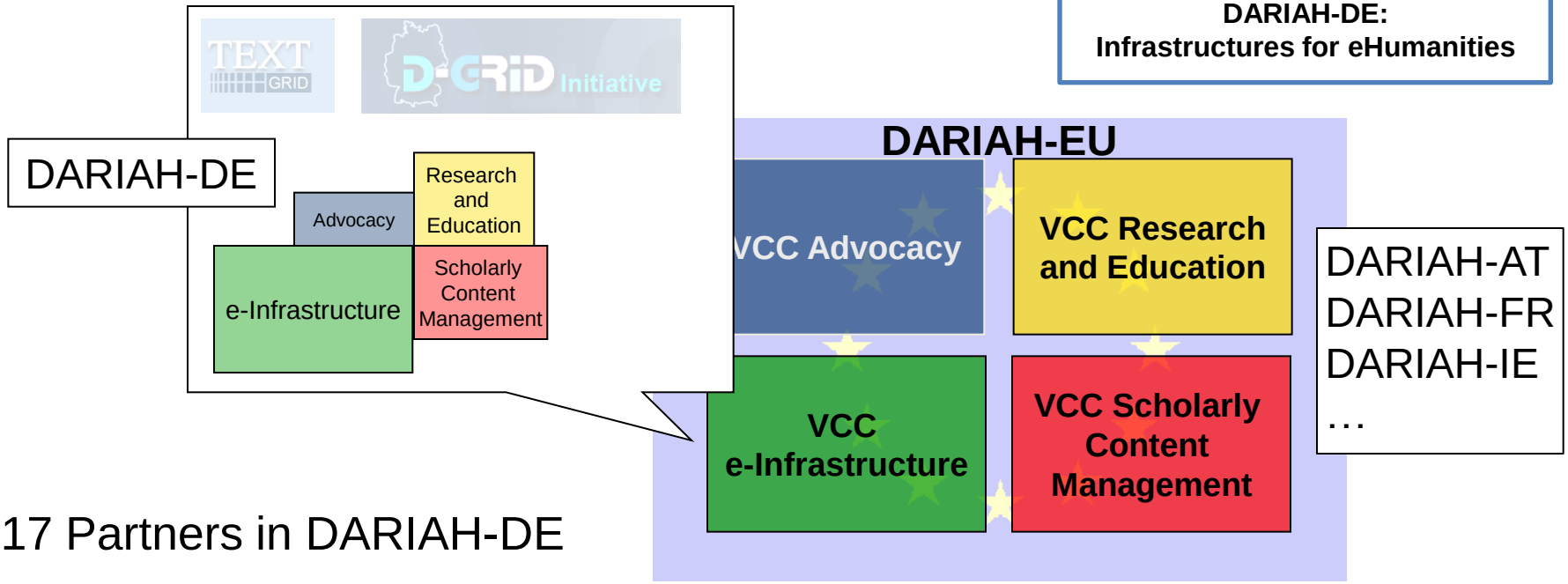
TECHNISCHE UNIVERSITÄT DARMSTADT

- **DARIAH-EU** network
- **D**igital **R**esearch **I**nfrastructures for the **A**rts and **H**umanities
- Sister project: **CLARIN-EU**

Bundesministerium für Bildung und Forschung

DARIAH
Digital Research Infrastructure for the Arts and Humanities

DARIAH-DE:
Infrastructures for eHumanities



- 17 Partners in DARIAH-DE

DARIAH-DE@UKP/TU Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT

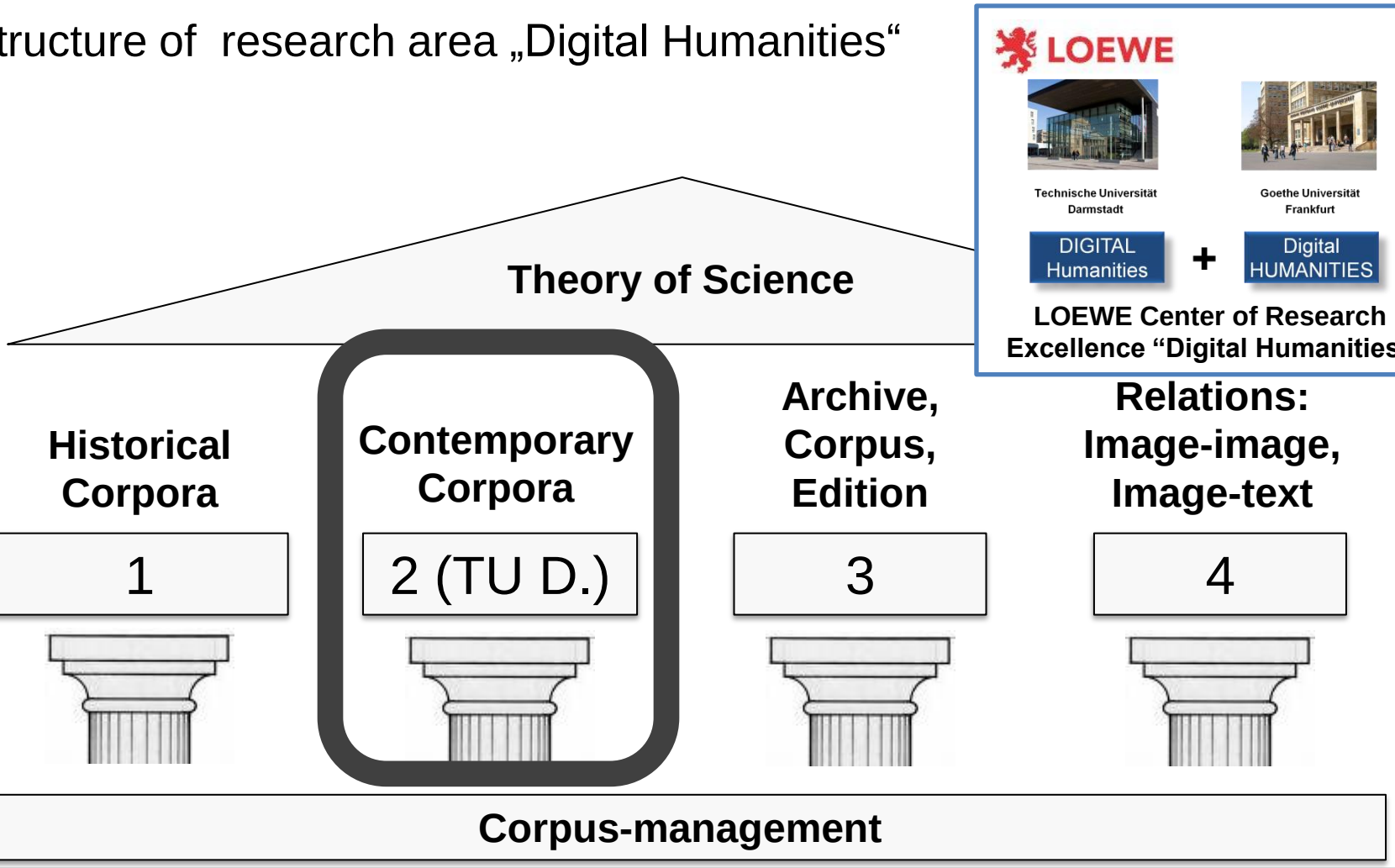
- **WP 1 E-Infrastructure**
  - Develop services for arts/humanities
    - User-oriented demonstrators
    - interactive community platforms
  - Focus on interoperability of data and services

- **WP 2 Research and Education**
  - Guide paradigm shift to information technology based research methods
  - Develop
    - Concepts for curricula for the eHumanities
    - Discipline-specific virtual research environments

LOEWE research area „Digital Humanities“

TECHNISCHE UNIVERSITÄT DARMSTADT

- Structure of research area „Digital Humanities“



UKP

„Digital Humanities“@UKP/TU Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT

2
Text als Instanz des Sprachsystems

Corpus-based analysis of Non-canonical grammatical constructions in English and German

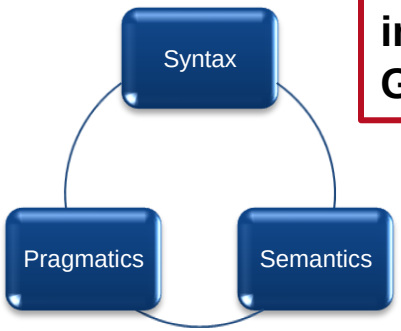
LOEWE

Technische Universität Darmstadt

Goethe Universität Frankfurt

DIGITAL Humanities + Digital HUMANITIES

LOEWE Center of Research Excellence “Digital Humanities”



3
Text als Prozess

Collaborative text development: Wikipedia Revision History

1
Text als Produkt

Machine Learning for Corpus Analysis

Language Technology (LT) for eHumanities

TECHNISCHE UNIVERSITÄT DARMSTADT

METHODS

LT for eHumanities: Tools and Methods

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Tools**: **DKPro component repository**
  - Java-based toolkit using the Apache UIMA framework
  - Process „flawed text“, e.g. not orthographically normalized text

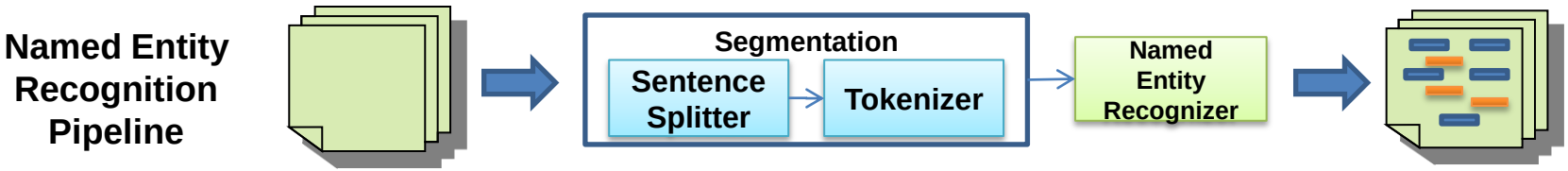
- **Method: Mining Multiword Expressions from Wikipedi**a
  - Multiword Expressions (MWEs):
    - Significant part of natural language
    - Many types are productive
    - Important for corpus linguistics and natural language processing of modern language
    - Relation to formulaic expressions

DKPro I

Darmstadt **K**nowledge **Pro**cessing Repository

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Highlights**
  - Conventiently compose natural language processing (NLP) pipelines
  - Reusable software components for NLP
  - General-purpose components for preprocessing
  - Special-purpose components for specific domains and problems
  - Based on Apache UIMA (Unstructured Information Management Architecture)



- **Applications at UKP**
  - Semantic text similarity
  - Key-phrase extraction
  - Opinion mining
  - Information Retrieval
  - Self organizing Wikis

http://www.ukp.tu-darmstadt.de/research/projects/dkpro/

DKPro II

Darmstadt **K**nowledge **Pro**cessing Repository

TECHNISCHE UNIVERSITÄT DARMSTADT

- **DKPro Core**
  - Linguistic **Preprocessing**
    - Tokenization, Compound splitting, POS-tagging, Parsing
  - Support for various file formats and data sources
    - XML, PDF, WSDL, Wikipedia
  - Available as open source: **http://dkpro-core-asl.googlecode.com**
- **DKPro IR**
  - Text indexing and retrieval
  - Support for classical and semantic methods
- **DKPro Semantics**
  - Keyphrase extraction, text segmentation, summarization, semantic relatedness, link discovery

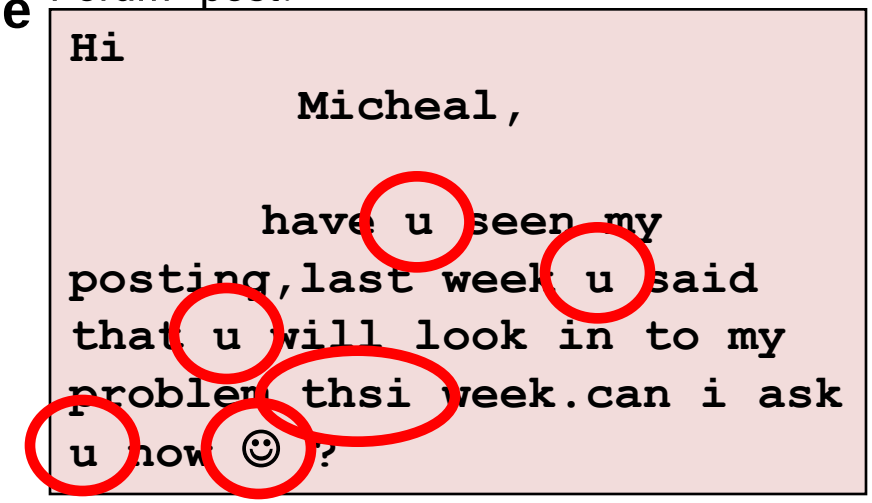
DKPro III

Darmstadt **K**nowledge **Pro**cessing Repository

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Use-case: user generated discourse**
  - NLP tools choke on
    - *Emoticons*
    - *Shorthand*
    - *Spelling errors*
    - *Spurious formatting/markup*
    - *Incorrect encoding conversion artifacts*
    - *…*
  - **Data Cleansing**
    - *Spell Checking, Spelling Correction*
    - *Dictionary Annotator*
    - *Regular Expression Annotator/Stemmer*
    - *Apply Changes Annotator*

Forum post:

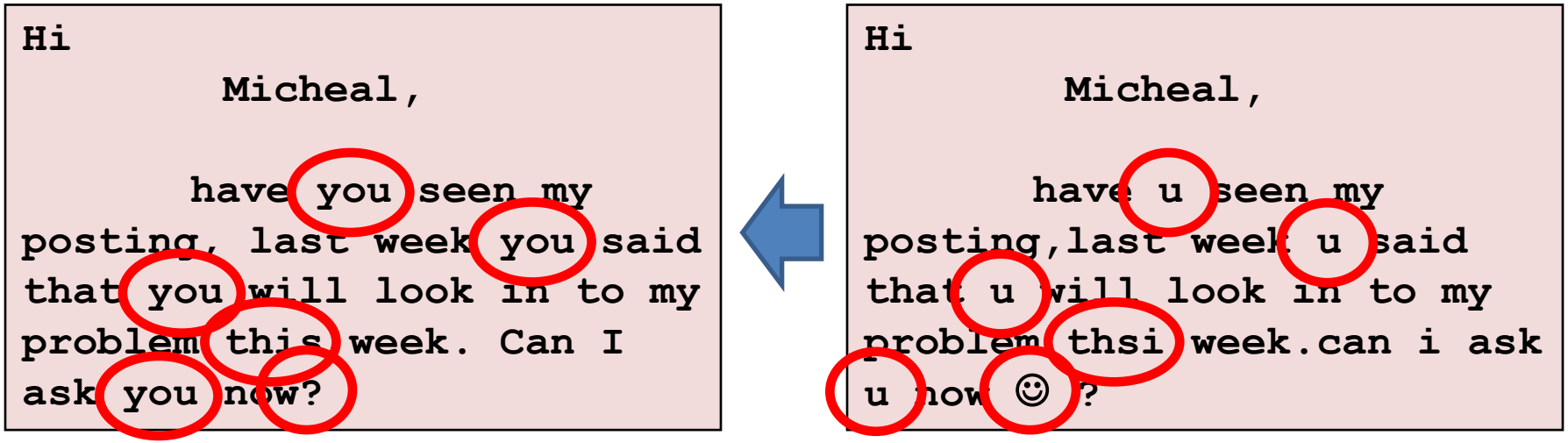


UKP

DKPro IV

Darmstadt **K**nowledge **Pro**cessing Repository

TECHNISCHE UNIVERSITÄT DARMSTADT



- Removed emoticons
- Expanded shorthand
- Corrected spelling errors

- *Analogy: Similar "flaws" in manuscripts*

TECHNISCHE UNIVERSITÄT DARMSTADT

Application Scenario: Mining Multiword Expressions from Wikipedia

LT for eHumanities: Multiword Expressions

TECHNISCHE UNIVERSITÄT DARMSTADT

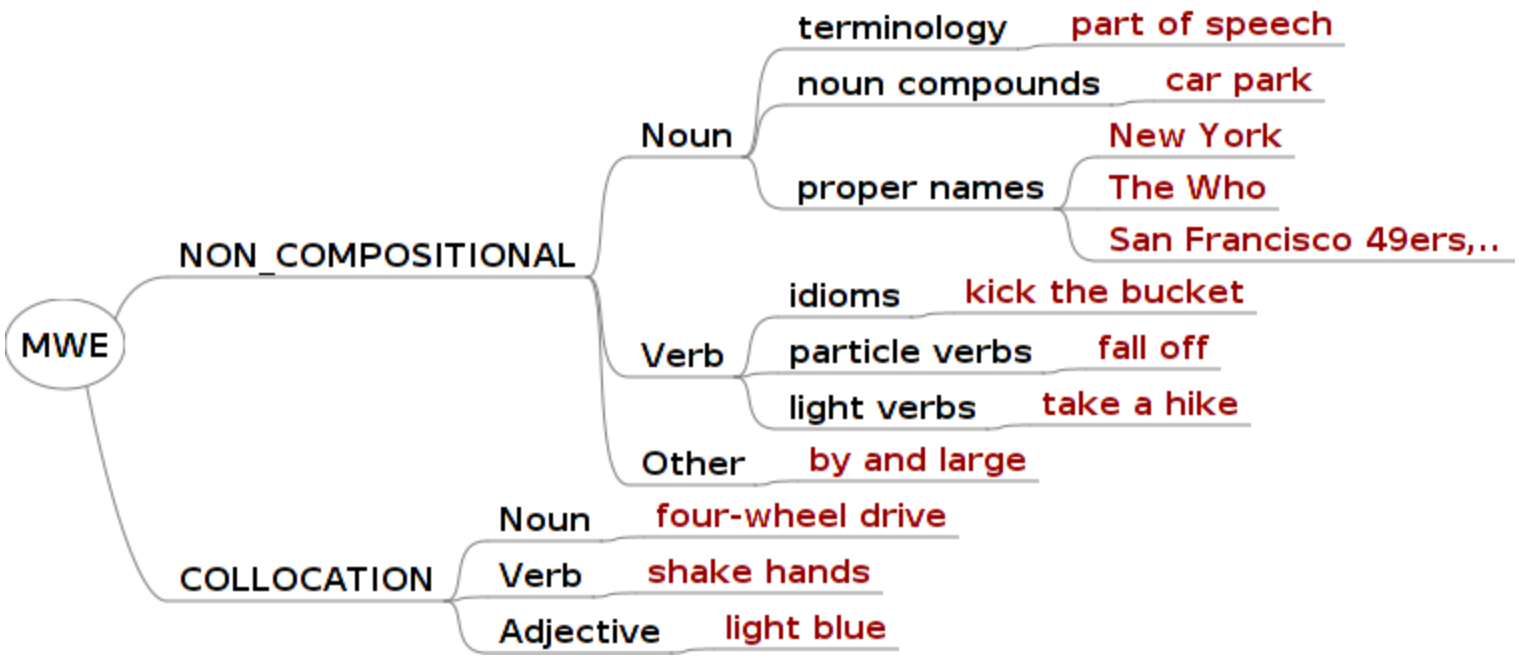
- Example: Mining English Multiword Expressions (MWEs) from Wikipedia
  - **MWEs**: „*Idiosyncratic interpretations that cross word boundaries*“ (Sag et al. 2002)
  - Statistically, syntactically or semantically irregular

  - **Motivation**
    - MWEs should not by split by parsers, during indexing
    - MWEs occur frequently, are productive
    - Application in Information Retrieval, Question Answering, Machine Translation, Ontology Creation,…
  - **Approach**
    - Corpus-based approach
      - Identification of candidates
      - Ranking (statistical association metrics)
      - Selection
    - Exploit structural properties of Wikipedia text („wisdom of the crowds“)

Mining Multiword Expressions

TECHNISCHE UNIVERSITÄT DARMSTADT

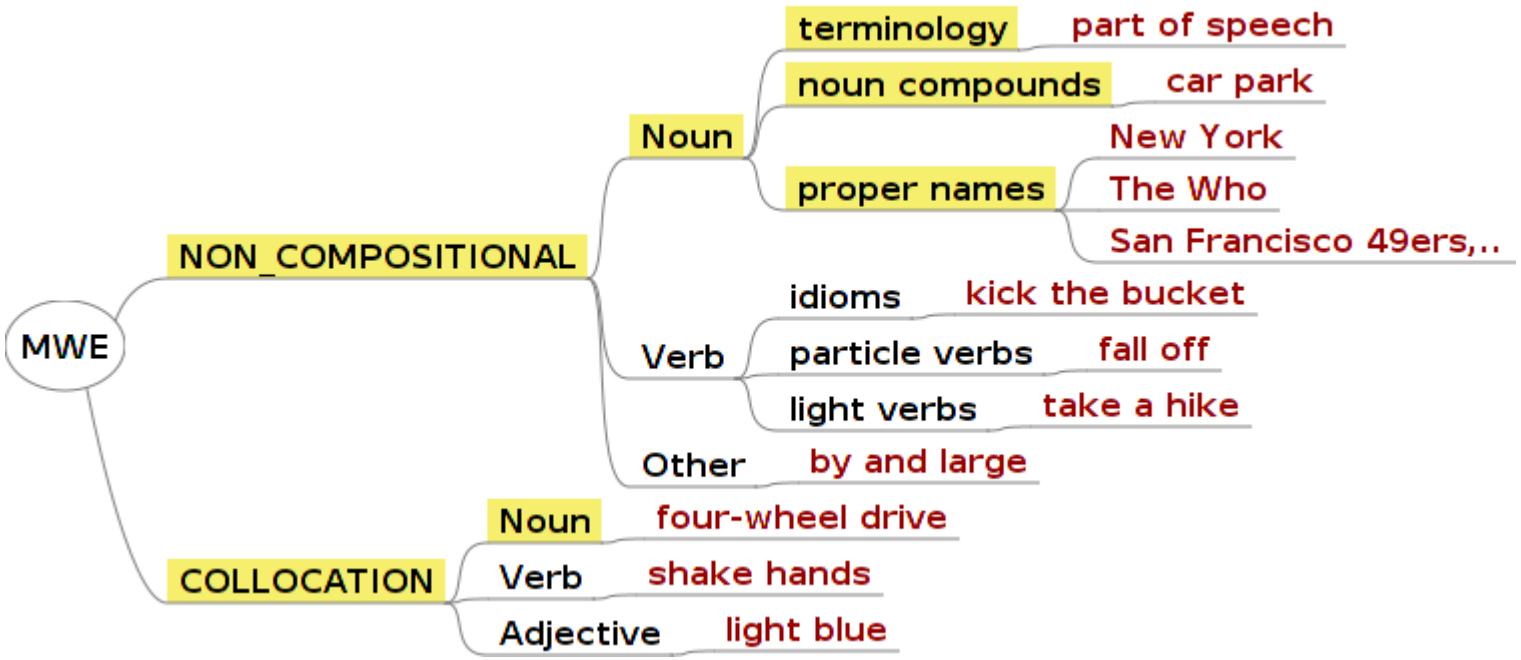
- **MWEs**: „*Idiosyncratic interpretations that cross word boundaries*“
- Statistically, syntactically or semantically irregular
- Example classification:



UKP

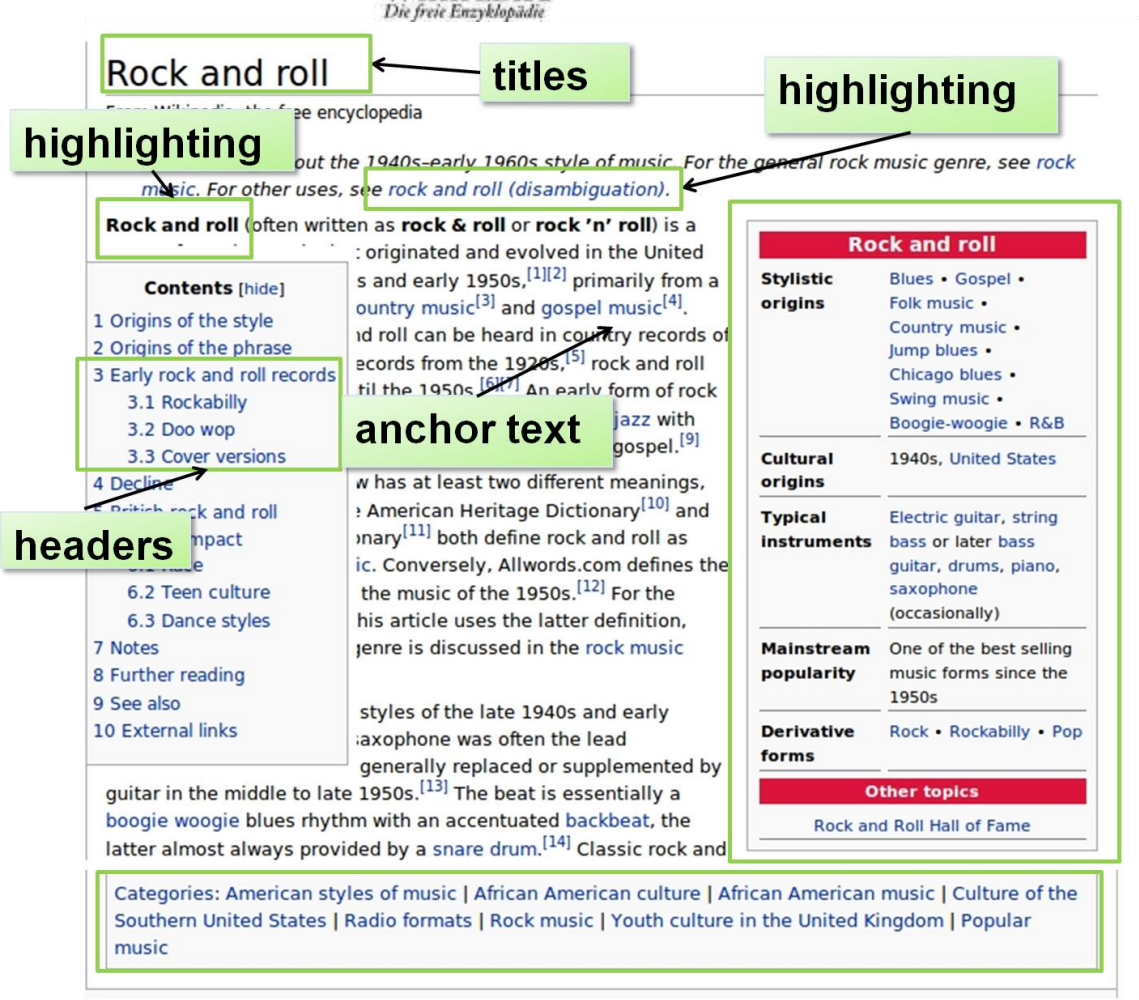
Mining Multiword Expressions

- **MWEs**: „*Idiosyncratic interpretations that cross word boundaries*“
- Statistically, syntactically or semantically irregular
- From Wikipedia (WikiMwe):



MWE from Wikipedia (WikiMwe)

- Candidate sources:
  - Marked-up sequences
  - Meaningful units
  - Benefit from user effort



JWPL

Java-based **W**ikipedia **P**rogramming **L**ibrary

TECHNISCHE UNIVERSITÄT DARMSTADT

Highlights

- High performance access to **Wikipedia content**
- Parser for the **WikiMedia syntax**
- Articles, discussion pages and categories as Java objects
- Access to information nuggets
  - **Redirects**, **links**, **I**
  - **ink anchors**, **interlanguage links,**
  - **sections, first paragraph, etc**.
- Supports all Wikipedia **language editions**

➡ low resource languages
➡ transfer from languages with better resources

Available open source: http://jwpl.googlecode.com

Reference

Torsten Zesch, Christof Müller and Iryna Gurevych (2008).
Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In: Proceedings of the Conference on Language Resources and Evaluation (LREC'08), electronic proceedings.

JWPL

Java-based **W**ikipedia **P**rogramming **L**ibrary

TECHNISCHE UNIVERSITÄT DARMSTADT

JWPL TimeMachine

- Tool for reconstructing past states of Wikipedia
  - Define a point in time and revert Wikipedia to that date
  - Makes reproduction of Wikipedia-based research results easy

JWPL RevisionMachine

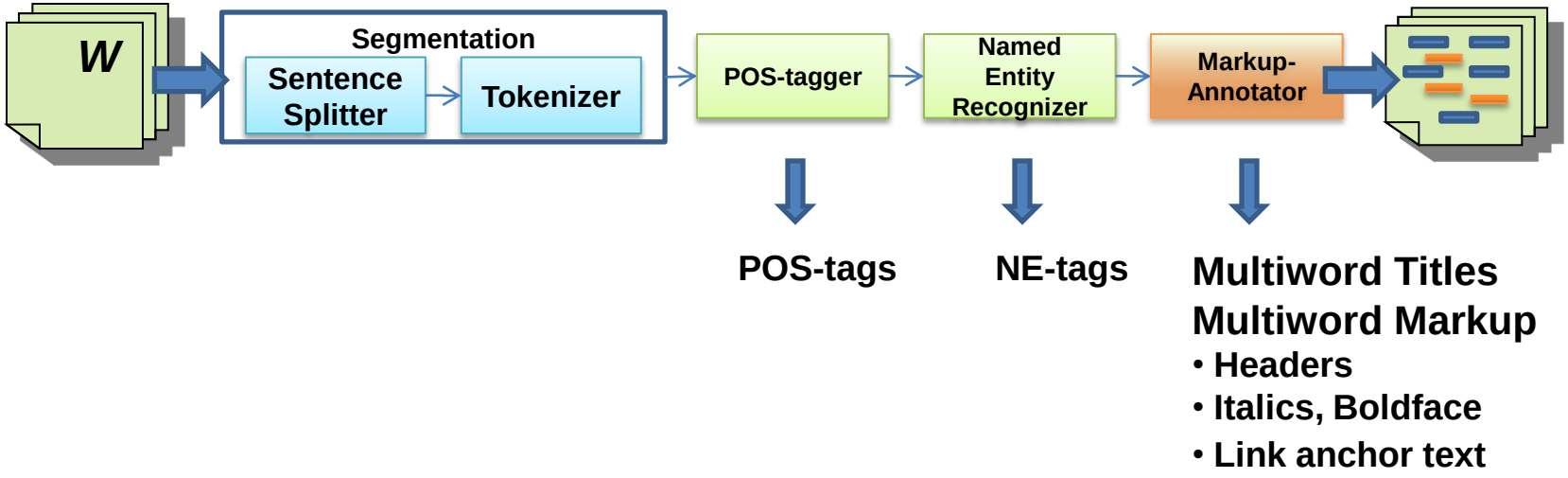
- Tool for easy access to **article revisions**
- Iterate over **all changes** to any article in Wikipedia
- Dedicated diff format for efficient storage of revision data
  - Complete uncompressed dump of a current English Wikipedia: ~5600 GB
  - Complete **JWPL database** of English Wikipedia with revisions: ~ 66 GB

⇨ study text development
⇨ study collaborative writing processes, interaction

Soon to be available open source

WikiMwe Candidate Extraction

- MWE candidate extraction pipeline:



  - Segmentation
  - POS-tagging
  - Named Entity Tagging
  - Markup Annotation
  - Output: candidate list (> 5 million candidates, 1.6 from titles, 4.3 from markup)
    - Later restrictions on candidates consisting of 2 to 4 constituent words

WikiMwe Candidate Ranking and Selection

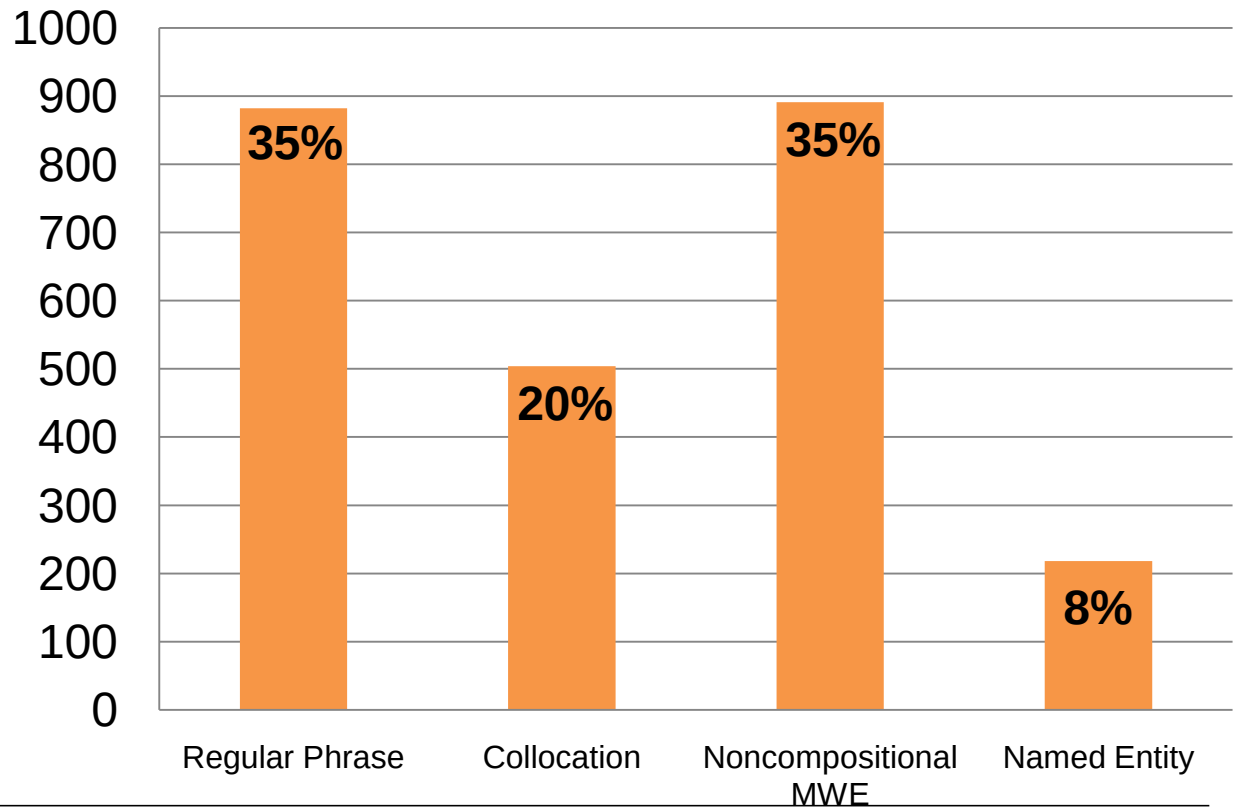
TECHNISCHE UNIVERSITÄT DARMSTADT

- Extraction of frequencies for candidates from Wikipedia text corpus
- Candidate ranking using statistical association metrics
- Candidate selection based on cutoff value
- Additional Filtering:
  - Wikipedia-typical patterns („List of Countries“)
  - Separation of Named Entities (NEs) based on
    - Stanford NE-tags
    - POS-sequences

- Numbers:
  - **> 500,000 WikiMwe-NEs**
  - **> 350,000 WikiMwes**

WikiMwe Manual Evaluation

- Annotation study on a random sample of **2500 WikiMwes**
- 2 Raters (linguist background), disambiguation to Gold Standard by expert
- Distribution of classes in the Gold Standard:
  - **> 55% „true MWEs“**



WikiMwe Examples

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Noncompositional MWE**
  - Upper triangular matrix
  - Oxygen mask
  - Choir screen
- **Collocation**
  - Low-cost airline
  - Burning sensations
  - First-person narrator
- **Regular Phrase**
  - Collaboration with Israel
  - Protest against the war
  - Camera models
- **Named Entities**
  - American Screenwriters Association
  - Breslau Seminary
  - Treatise on rhetoric

WikiMwe Evaluation: Comparison

- **WikiMwe resource:**
  - ca 350,000 (noisy) MWEs of size 2-4
  - Estimate based on annotated sample: **>190,000 „true MWEs“**

- Other MWE resources:
  - **MWE data sets**: Small (often manually crafted) evaluation sets
    - Noun Compound by C. Bannard: ~ 500
    - Noun Compounds by Tratz & Hovy 2010: 17,000
  - Specialized dictionaries: idioms, etc.

  - **WordNet**: >68,000 MWEs, 63,000 noun compounds of size 2-4
    - covered > 70% by WikiMwe
  - **Wiktionary**: > 40,000 MWEs

(…and this is for English..)

Summary WikiMwe

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Conclusion**
  - Identification of multiword expressions is still a big problem for language processing
  - We leverage Wikipedia annotations to help with this task
  - Result:
    - WikiMwe resource: large, noisy set of MWEs and NEs
    - Mainly nouns
    - Strong in technical terminology
    - Domain-transcending
  - **Potential applications for WikiMwe:**
    - IR/QA (few harm by noise)
    - Ontology creation (technical terminology)
  - **Open questions:**
    - How to disambiguate MWEs (type vs token): *plain dress*, *red carpet*, *don't drink the water*?
    - Wikipedia as a corpus for collocation extraction (compared to newswire)
    - Transfer to other languages

LT for eHumanities: Summary

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Flexible Tools:**
  - DKPro processing pipeline
  - Data cleansing
  - Additional analysis components as needed

- **Example application:** Mining MWEs from Wikipedia
  - Exploit structural properties of available data
  - Efficiently process textual data
  - Apply to languages with few resources

- *Analogy: User generated „flawed text“ and historical documents*

➡ Future goal: Tools adapted to eHumanities (*DKPro eHum*)

Thank you for your attention!

TECHNISCHE UNIVERSITÄT DARMSTADT

KTS
KLAUS TSCHIRA STIFTUNG GEMEINNÜTZIGE GMBH

GRADUIERTENKOLLEG E-LEARNING

Bundesministerium für Wirtschaft und Technologie

VolkswagenStiftung

Deutsche Forschungsgemeinschaft DFG

Emmy Noether-Programm
Deutsche Forschungsgemeinschaft DFG

http://www.ukp.tu-darmstadt.de