CLARIN Annual Conference 2014 in Soesterberg, the Netherlands

Thomas Bartz & Christian Pölitz

Using Data Mining and the CLARIN Infrastructure to Extend Corpus-Based Linguistic Research

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Corpus-Based Linguistic Research and Analysis Using Data Mining

TU Dortmund University
Department of German Language and Literature & Artificial Intelligence Group

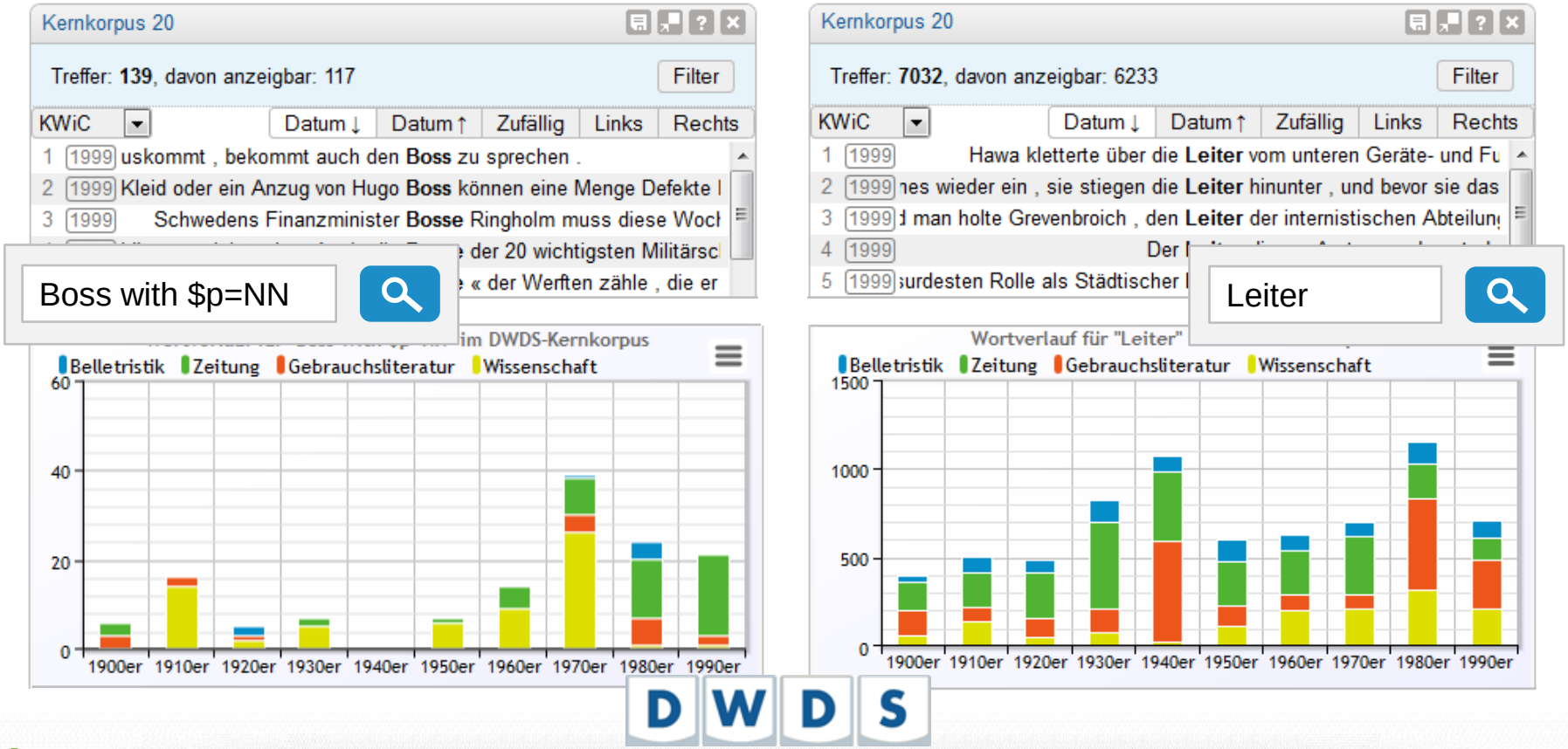
Using Data Mining and the CLARIN Infrastructure to Extend Corpus-Based Linguistic Research

1. Motivation
2. Project Background
3. Application and Evaluation
4. Conclusion and Outlook

For Example

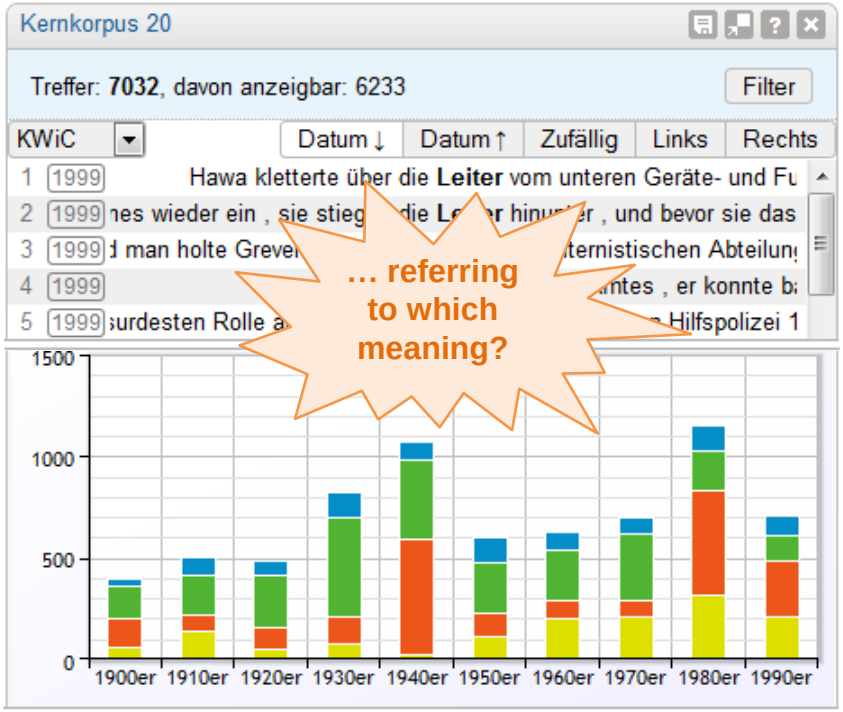
Do Borrowings Like *Boss* Replace Indigenous German Words?

Large digital corpora of written language offer **excellent possibilities for linguistic research on authentic language data**.



Large digital corpora of written language offer excellent possibilities for linguistic research on authentic language data.

But if the queried word forms or constructions are (semantically) ambiguous, intense **manual effort has to be done for cleaning and disambiguation tasks** (depending on the amount of data).



Leiter has different meanings:

1. chief, director
2. ladder
3. conducting medium
4. scale (music)

Researchers actually aren‘t aware of the different meanings the KWIC snippets and word occurrence statistics refer to.

Corpus-Based Linguistic Research
Cleaning and Analyzing Tasks

- **Filtering:** KWIC lists are not immediately useful because they contain many **false positives** that cannot be filtered out with currently available corpus technology.
- **Disambiguation and visualization:** A word whose use is under investigation exists with **various meanings**, which in the context of the investigation **need to be differentiated**. This is particularly crucial with regard to visualisations of word frequencies.
- **Annotation and classification:** A linguistic phenomenon cannot be accurately pinpointed with the help of an available corpus query language or alternatively a stage of analysis necessitates **finer** (post-hoc) **classification of the collected data** (e.g. via efficient task-specific annotation).
- **Detection of interesting results:** A large number of examples for a frequent linguistic pattern can be found that are of a similar type, while there are some **seldom-used, but very interesting examples** (e.g. metaphors) that are sought.

The KobRA Project
Corpus-Based Research and Analysis Using Data Mining

Project Partners

technische universität dortmund

UNIVERSITÄT MANNHEIM

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

INSTITUT FÜR DEUTSCHE SPRACHE

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Linguistics

Computer Science

Computational Linguistics
Language Ressources / CLARIN

Aims

- Improvement and acceleration of quantitative analysis of structured language data
- **Customization and evaluation of data mining techniques** (machine learning in particular) in the context of corpus-based linguistic studies in the fields of:
  - Diachronic linguistics
  - Corpus-based lexicography
  - Variational linguistics

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

KobRA

The KobRA Project
Language Resources and Data Mining Techniques

Important Language Resources

Corpus	Data	Size	Annotations
DWDS Core Corpus of the 20th century	balanced over time and by text genre	100M tokens	lemmas, PoS tags, metadata
ZEIT Corpus	newspaper articles (1946-2014)	460 M tokens	lemmas, PoS tags, metadata
German Text Archive	balanced over time (1600-1900) and by text genre	100M tokens	lemmas, PoS tags, metadata
German Reference Corpus	different text genres (1900-2013)	24B tokens	lemmas, PoS tags, metadata
Wikipedia Corpus	article and talk pages (2013)	1B tokens	lemmas, PoS tags, metadata
Tübingen Treebank of Written German	newspaper articles (German newspaper “taz – die tageszeitung”)	1.5M tokens	lemmas, PoS tags, morphology, syntax, coreferences, named entities

Berlin-Brandenburg Academy of Sciences and Humanities

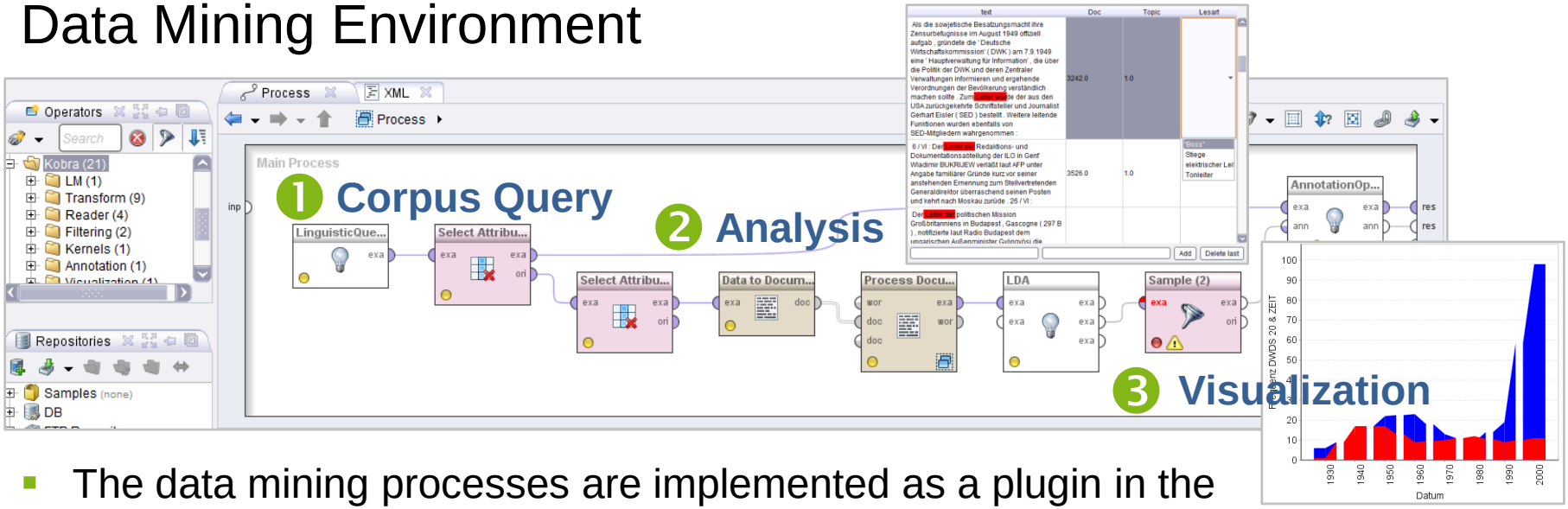
Institute for the German Language

Tübingen University, Department of Computational Linguistics

Data Mining Techniques

- Operating on search result KWIC snippets
  - Extraction of linguistic and document features
- Supervised classification: Support Vector Machines
  (cf. Joachims, 1998, 2002; Cristianini & Shawe-Taylor, 2004)
- Unsupervised disambiguation: Topic Models
  (LDA, cf. Blei et al., 2003; Steyvers et al., 2004; Blei & Lafferty, 2006)

The KobRA Project
Data Mining Environment



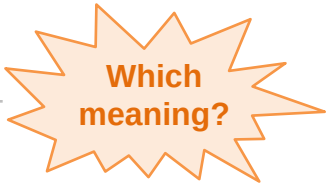
- The data mining processes are implemented as a plugin in the data mining framework RapidMiner (formerly: 'YALE', Mierswa et al., 2006)
  - This allows one to perform large scale data analysis and
  - offers methods to import, transform, analyze and visualize data
- The KobRA plugin provides:
  - Direct access to language resources
  - Methods to extract linguistic and document features from KWIC lists
  - Methods for classification and disambiguation
  - An integrated annotation environment

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Case Studies in the Field of Corpus-Based Lexicography
Do Borrowings Like *Boss* Replace Indigenous German Words?

❶ Data Retrieval DWDS Core Corpus of the 21st Century

Query	Occurrences
”Boss with $p=NN“ boss, chief, director; named entity	139
”Leiter“ boss, chief, director; ladder; conducting medium; scale (music)	7032
”Vorgesetzter with $p=NN“ boss, chief, director	1807
”Chef“ boss, chief, director	5908



❷ Unsupervised Disambiguation

- Source: KWIC query result lists; word context: 3 sentences
- Method: Latent Dirichlet Allocation (LDA); representation: bags of words
  (cf. Blei et al., 2003; Steyvers et al., 2004; Blei & Lafferty, 2006)

❷ Unsupervised Disambiguation

- Hypothesis: **Certain distributions of words** in each KWIC snippet and over the KWIC query result list **correspond to certain meanings of a queried word**. For example *Leiter* with the meaning “ladder”: *steigen* (climb), *auf* (upon), *hohe* (high) etc.
- **LDA models the probability distributions of the words and the KWIC snippets** from the result list. The probability distributions are scattered over a number of “latent topics” that correspond to different meanings of the queried word.
- The estimation of the distributions is done via a Gibbs sampler (Griffiths & Steyvers, 2004). The **Gibbs sampler models the process of assigning a word or snippet to a certain topic** based on the topic distributions of all other topics.
- We investigate the possibility to **integrate further information into the generation of the topic models**: We use the approach of Steyvers et al. (2004) to integrate information about text genre classes. Moreover, we use dynamic topic models (cf. Blei & Lafferty, 2006) that facilitate analyzing the development of topics over time.

3 Evaluation

Manual Disambiguation of 30% of the Retrieved Data

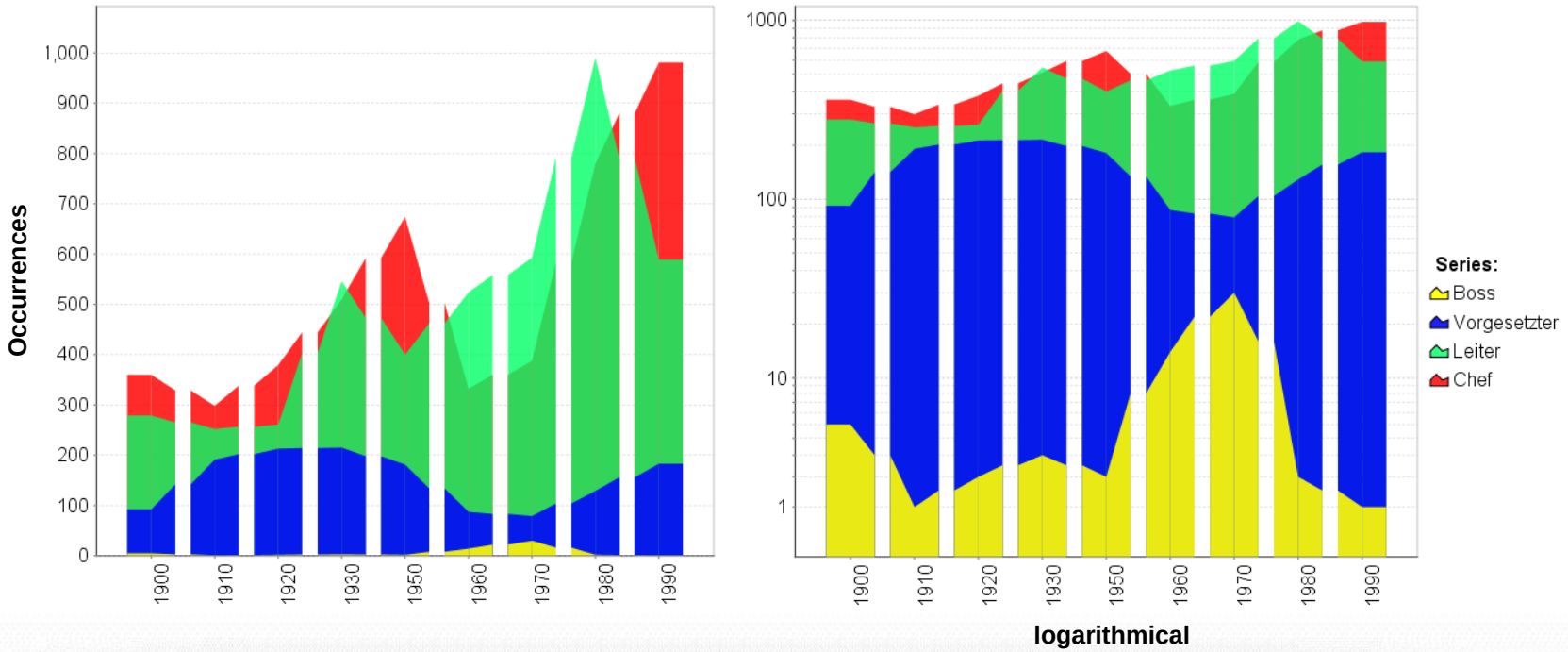
Leiter	Boss, chief, …	Ladder
kappa	0,97	
F_1	0,97	0,85
NMI	0,300	

Boss	Boss, chief, …	NE
kappa	0,98	
F_1	0,88	0,87
NMI	0,571	

2 Annotators

LDA

4 Visualization

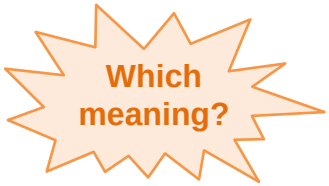


Case Studies in the Field of Corpus-Based Lexicography
Tracing Semantic Development

❶ Data Retrieval

DWDS Core Corpus of the 21st Century[1], DWDS ZEIT Corpus[1], Leipzig Corpora Collection[2] (cf. Quasthoff, Richter and Biemann, 2006)

Query	Occurrences
”zeitnah“[1] contemporary, critical of the times vs. prompt	592
”cloud“[2] mass of condensed water, smoke, dust or other elements vs. named entity vs. remote server network	1486



❷ Unsupervised Disambiguation

- Source: KWIC query result lists; word context: 3 sentences[1] / 1 sentence[2]
- Method: Latent Dirichlet Allocation (LDA); representation: bags of words
  (cf. Blei et al., 2003; Steyvers et al., 2004; Blei & Lafferty, 2006)

❸ Evaluation Manual Disambiguation of 30% of the Retrieved Data

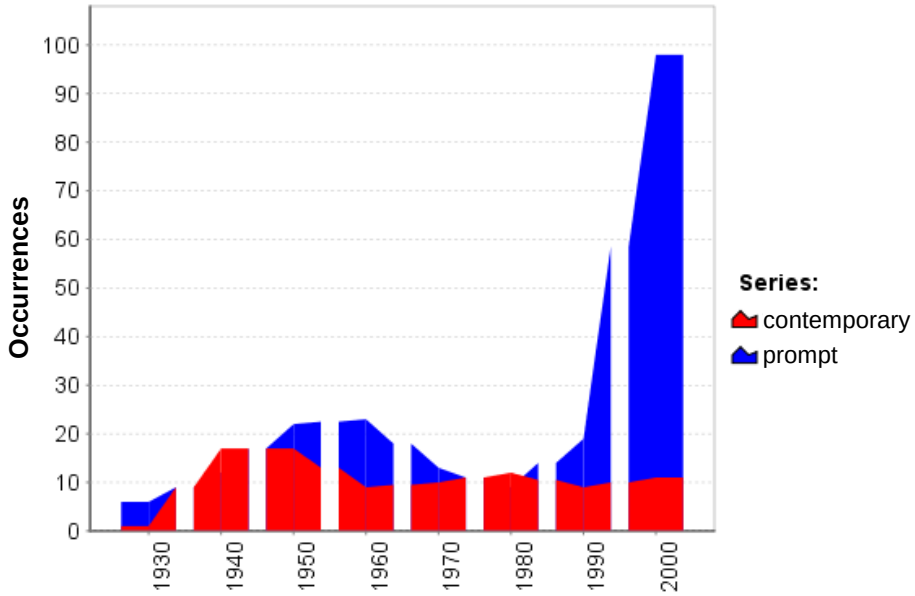
zeitnah	contemp., …	prompt
kappa	0,91	
F_1	0,88	0,88
NMI	0,418	

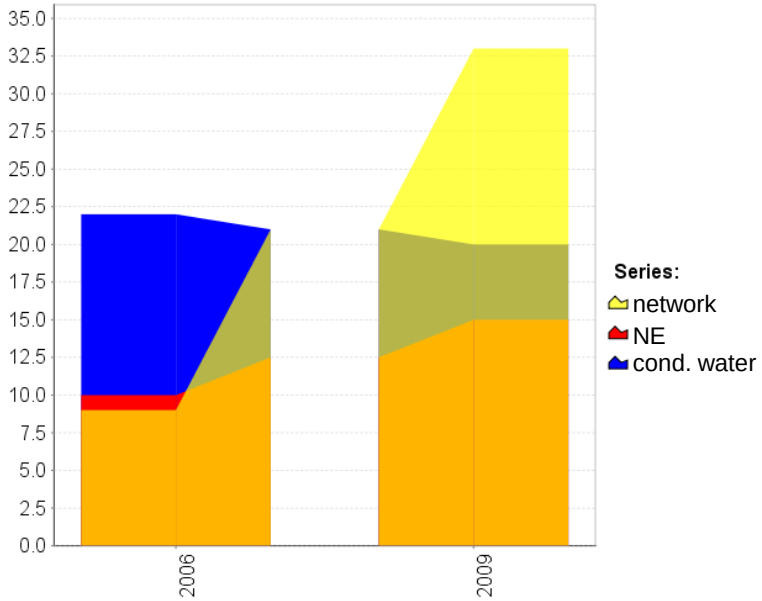
cloud	cond. water, …	NE	network
kappa	0,92		
F_1	0,85	0,81	0,63
NMI	0,366		

2 Annotators

LDA

❹ Visualization





Conclusion and Outlook

- We introduced the approach and selected results of the KobRA project, that is directed to **improve and accelerate corpus-based linguistic analysis using the CLARIN infrastructure**.
- To this end, we deploy and **evaluate data mining techniques for the preparation and structuring of corpus query result KWIC lists**.
- **Case studies** from the field of corpus-based lexicography already **illustrate the benefit** of the approach.
- Currently we are evaluating and, as needed, adapting the processes using **further examples and data from additional (particularly diachronic) corpora**. Thereby, we are also testing the consideration of additional linguistic features (PoS, dependency, etc.) in machine learning.
- At the end of the project term, the evaluated processes will be made available in a data mining environment for corpus-based linguistic research and further fields of study in the digital humanities. They will also be **integrated into the CLARIN infrastructure**.

Thank you!

References

Bartz, T., Pölitz, C. and Radtke, N. (2013). *Automatische Klassifikation von Stützverbgefügen mithilfe von Data-Mining*. Technischer Bericht, Technische Universität Dortmund. http://kobra.tu-dortmund.de/mediawiki/images/a/a1/KobRA-MS1a_Belegklassifikation.pdf (accessed 01 October 2014).

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation, *Journal of Machine Learning Research,* 3: pp. 993–1022.

Blei, D.M. and Lafferty, J.D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. New York: ACM, pp. 113–120.

Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20 (1): 37–46.

Cristianini, N., and Shawe-Taylor, J. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.

Griffiths, T.L. and Steyvers, M. (2004). Finding scientific topics, *Proceedings of the National Academy of Sciences* 101 (suppl 1): 5228–5235.

Joachims, T. (1998): Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the European Conference on Machine Learning*. Berlin, Heidelberg: Springer.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines*. Dissertation, Dordrecht: Kluwer.

Lüdeling, A. and Kytö, M. (eds) (2008/9). *Corpus Linguistics. An International Handbook.* 2 Bände. Berlin, New York: de Gruyter.

Manning, C.D., Raghavan, P. and Schütze, H. (2008). *Introduction to Information Retrieval*. New York: Cambridge University Press.

McEnery, T., Xiao, R. and Tono, Y. (2006). *Corpus-Based Language Studies. An Advanced Resource Book* (Routledge Applied Linguistics). London, New York: Routledge.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M. and Euler, T. (2006): YALE: Rapid Prototyping for Complex Data Mining Tasks. In *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining*. Philadelphia, USA.

Navigli, R. and Crisafulli, G. (2010). Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, pp. 116–126.

Quasthoff, U., Richter, M. and Biemann, C. (2006). Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the fifth international conference on Language Resources and Evaluation*. Genoa, pp. 1799–1802 .

Steyvers, M., Smyth, P., Rosen-Zvi, M. and Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York: ACM, pp. 306–315.

Storrer, A. (2011). Korpusgestützte Sprachanalyse in Lexikographie und Phraseologie. In Knapp, K. et al. (eds), *Angewandte Linguistik. Ein Lehrbuch*. Tübingen: Franke, pp. 216–239.