

# Data Mining in Digital Humanity

Dipl. Inform. Christian Pölitz

TU Dortmund

# Corpus-Based Research and Analysis

## Using Data Mining

### Project Partners



Linguistics

Computer  
Science

Computational Linguistics  
Language Resources / 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



# Corpus Linguistic Data Mining Plugin

The screenshot displays the RapidMiner 5.3.015 interface with the following components:

- Process Canvas:** A workflow diagram titled "Main Process" showing the following sequence of operators:
  - Date to Numerical:** Converts date attributes to numerical values.
  - Select Attributes:** Selects specific attributes for processing.
  - Nominal to Date:** Converts nominal date attributes to a standard format.
  - Select Attributes:** Another selection step for the processed data.
  - Process Documents:** Processes the selected data into document format.
  - Data to Documents:** Converts the processed data into a document structure.
  - LDA:** Performs Latent Dirichlet Allocation on the documents.
- Parameters Panel (LinguisticQueryOperator):**
  - gui.action.wizard.Query.label
  - query: Leiter
  - encapsulation: (empty)
  - sample size: 5000
  - encoding: UTF-8
  - data source: DWDS20
  - context size: 3
  - extract lemmas:
  - extract tags:
- Problem Log:** Shows two potential problems:
  - Message:** "The attribute 'date' is missing in the input example set." / "The attribute 'date' is missing in the input example set."
  - Fixes:** "Change value of parameter 'attri...' / "Change value of parameter 'attri...'"
  - Location:** "Nominal to Date.exempl..." / "Date to Numerical.exam..."

# Text Resources

- Text Corpora
  - Deutsches Textarchiv
  - Wörterbuch der deutschen Sprache
  - Dictionaries
  - Word profiles and statistics
  - Word correlations (Wordnet)

Corpus	Data	Size	Annotations	
<b>DWDS Core Corpus of the 20<sup>th</sup> century</b>	balanced over time and by text genre	100M tokens	lemmas, PoS tags, metadata	Berlin-Brandenburg Academy of Sciences and Humanities
<b>ZEIT Corpus</b>	newspaper articles (1946-2014)	460 M tokens	lemmas, PoS tags, metadata	
<b>German Text Archive</b>	balanced over time (1600-1900) and by text genre	100M tokens	lemmas, PoS tags, metadata	
<b>German Reference Corpus</b>	different text genres (1900-2013)	24B tokens	lemmas, PoS tags, metadata	Institute for the German Language
<b>Wikipedia Corpus</b>	article and talk pages (2013)	1B tokens	lemmas, PoS tags, metadata	
<b>Tübingen Treebank of Written German</b>	newspaper articles (German newspaper "taz – die tageszeitung")	1.5M tokens	lemmas, PoS tags, morphology, syntax, coreferences, named entities	Tübingen University, Department of Computational Linguistics

# Motivating example: Disambiguation / Topic Model

- Want to identify abstract topics in texts
- Texts are mixtures of topics
- Example:
  - Er kletterte die Leiter rauf. (Topic ladder)
  - Er wurde Leiter des Institutes. (Topic boss)

# Topic Models for Word Sense Induction

- Given a word, find possible senses
  - Based on co-occurrences
- Retrieve KWIC lists from text corpora
  - Key word in context
  - Additional meta information

Substantiv logDice 36

Überblick zu 'Kreatur'

**Achtung vor armselige Aufschrei**  
bedauernswerte bedrängten bemitleidenswerte  
bevölkern Elend elende erbärmliche gedemütigte gehetzte  
geknechtete gepeinigten gequälten  
**geschundene Harren hilflose Leid Leiden**  
leidenden Mitgefühl mit **Mitleid mit**  
**Mitleiden mit** Qual Schlüssel Schmerz Schrei  
Schöpfung **Seufzen Seufzer** Sinnbild unschuldige  
wehrlose Würde

Version: 3.0 Einstellungen

CREATUR, f.

geschöpf, aber tönender und mächtiger als das deutsche...

- 1) res creata: er wird seinen eifer...
- 2) der günstling und anhängen eines reichen,...
- 3) vorzugsweise gilt creatur ...
- 4) als schelte sagt man auch von einem...
- 5) die unredlichen creaturen = die läuse. weisth. 2,...

Version: @rev145 Kompakt | Details

KWIC Datum ↓ Datum ↑ Zufällig Links Rechts

- 1 [1903] tt bei dem Gedanken , welcher **Kreaturen** Sklavin Du bist !
- 2 [1903] erwecken , sich rückwärts von **Kreatur** zu **Kreatur** noch einmal a
- 3 [1900] i , in die große Sehnsucht aller **Kreatur** gegen Licht und Sonne ur
- 4 [1899] Hier ist von einem aller **Kreatur** angeborenen Erlösungsbe
- 5 [1899] rung der Hetzkapläne und ihrer **Kreaturen** ; im Guten , nicht im B
- 6 [1899] s wären sie Herren über Gottes **Kreaturen** , und frei von allen Ges
- 7 [1899] gt dem heiligen Hörer aus allen **Kreaturen** ; Alles was er ansieht
- 8 [1899] ist , damit wir arme , sinnliche **Kreaturen** ihn möchten fassen un
- 9 [1899] Verken ( von aussen gegen der **Kreatur** zu rechnen ) sind wir Chri:
- 10 [1899] Wiedergeburt " zu einer neuen **Kreatur** " , glaubt man , es sei Zuf
- 11 [1899] Herr , vor mir grault sich keine **Kreatur** . "
- 12 [1898] Jerbares Benehmen mit seinen **Creaturen** , ich habe die Anstellur
- 13 [1898] Selige **Kreatur** , sagt ein alter Grieche , -
- 14 [1898] der alles Leid und alle Lust der **Kreatur** auf seinen Schultern trägt
- 15 [1898] unsäglich selige Tod , den jede **Kreatur** in unendlicher brennender

Version: 1.0 Optionen

Kernkorpus 20

Treffer: 731, davon anzeigbar: 559

KWIC Datum ↓ Datum ↑ Zufällig Links Rechts

- 1 [1999] degenh: Es tagt , der Sonne Morgenstrahl weckt alle **Kreatur** .
- 2 [1999] duecke: seiner Gedankenlosigkeit oder mit seinem überlegten Griff eine **Kreatur** ausgewählt , die sehr viel mit ihm gemeinsam hat .
- 3 [1999] hars ernerwelt so sehr , daß die eigenen Väter Angst vor ihrer miesen **Kreatur** bekamen .
- 4 [1999] kurz acker , Bereicherungswütigen und » Erfolgsmenschen « . Diese **Kreaturen** des Marktes , die sich als Subjekte der » neuen Beweglich
- 5 [1999] kurz ungen mit ebenso düren wie klaren Sätzen bezeichnet : » Alle **Kreaturen** sind vom Tage ihrer Geburt an einsam und bedürfen einande
- 6 [1999] kurz Wozu dient es , frage ich , daß man solche **Kreaturen** mit soviel Mühe am Leben erhält ?
- 7 [1999] kurz rgerecht . . . Es gibt besonders unter den Frauen schreckliche **Kreaturen** . . . ( zit. nach : Perrot 1981 , 82 ) . Auch für das England
- 8 [1999] kurz ie kolossale Grundstücksspekulation und die Begünstigung der **Kreaturen** des Kaiserreichs . . . durch Zuwendung ungezählter Millione
- 9 [1999] kurz Es ist der Jammer der Welt , es ist die gemartete **Kreatur** , ein wilder , grauenvoller Schmerz , der da stöhnt .
- 10 [1999] moers r Geräusche und grusliger Gesänge geme möglichst wehrlosen **Kreaturen** , um sich an deren Unbehagen zu ergötzen .
- 11 [1999] moers te Blaubärchen der Welt war , die absolut bemitleidenswerteste **Kreatur** , die jemals . . . und endlich flossen die Tränen !
- 12 [1999] moers Alles in allem zeigten diese ansonsten gefühllosen **Kreaturen** ein erstaunliches Maß an Begeisterung .
- 13 [1999] moers lerglutze , und Berghutzen sind wahrscheinlich die häßlichsten **Kreaturen** , die man sich vorstellen kann .
- 14 [1999] moers che ins Gesicht : Berghutzen sind mit Abstand die häßlichsten **Kreaturen** , die man sich vorstellen kann .
- 15 [1999] moers Mit so einer **Kreatur** am Hals würde es sicher nicht einfacher sein , es im wirklicher

Version: 1.1 Optionen

DIE ZEIT

Treffer: 1809

KWIC Datum ↓ Datum ↑ Zufällig Links Rechts

- 1 [2014] , Gott und Mensch , Held und **Kreatur** fallen widersprüchlich in e
- 2 [2014] es kein Problem für die kleinen **Kreaturen** .
- 3 [2014] rde ist zu erkennen , wie diese **Kreatur** entstand , wie der Menscl
- 4 [2014] Freunde , nahezu menschliche **Kreaturen** geworden , ausgestatte
- 5 [2014] ren haben sich aus der Urzelle **Kreaturen** wie der Farn , die Nack
- 6 [2014] Weshalb unterdrückt eine **Kreatur** die andere ?
- 7 [2014] Psychologie : Den " **Kreaturen** der Nacht " haben die f
- 8 [2014] chten muss , selbst von seiner **Kreatur** verschlungen zu werden .
- 9 [2014] :kkehren werde , laufen gerade **Kreaturen** durch die Straßen , die
- 10 [2014] t geschaffen , nicht von seinen **Kreaturen** .
- 11 [2014] , ist , dass Gottvater alle seine **Kreaturen** kennt und ihm nichts ve
- 12 [2014] ertiefe Oktopoden oder andere **Kreaturen** zu sehen .
- 13 [2014] Das heißt , es gibt womöglich **Kreaturen** da unten , die wir niem:
- 14 [2014] dschihadisten verachtenswerte **Kreaturen** .
- 15 [2014] mühen sich schon heute , ihre **Kreaturen** zu vermenschlichen .

Version: 2.1 Optionen für die Verbindung mit ITMC-WPA2 erforderlich  
Klicken Sie hier, um weitere Informationen anzugeben.

# Latent Dirichlet Allocation

- Generative model
- Words in texts are random variables
- Word emission depends on unobserved factors/topics

# Gibbs Sampler

- Idea:
  - For a given word in a document sample topic
  - Topic depends on all other topics but itself

# Integrating External Information

- Dictionaries
- Wordnets
- Historical information

### DWDS-Wörterbuch

**Kreatur** Aussprache: ►  
fem., -, -en  
Herkunft: Latein

- 1 Lebewesen, Geschöpf**  
*eine arme, bedürftige, gemarterte, hilflose Kreatur*
- 2 abwertend verachtenswerter Mensch**  
*eine armselige, bedenkenlose, gemeine, jämmerliche, nichtswürdige, undankbare, verkommene Kreatur*

**Günstling, Lakai**  
*Diese Kreaturen haben die Posten und Pöstchen, die ihrer Lakaienseele am besten zusagen – Traven General 127*

Version: 0.4.23 | Quelle: WDG | Artikeltyp: Vollartikel Kompakt | Details

### Etymologisches Wörterbuch

**Kreatur**

**Kreatur** f. 'Geschöpf, die geschaffene Natur', in abschätzigem Sinne 'gefügiger Mensch, willenloses Werkzeug in den Händen von Auftraggebern', mhd. *crēatūr(e)* 'Geschöpf' ist eine direkte Entlehnung aus kirchenlat. *creātūra* 'Schöpfung, Geschöpf' (zu lat. *creāre* '(er)schaffen, (er)zeugen'), während mhd. *crēature* auf dem (ebenfalls aus dem lat. Substantiv entlehnten) afrz. *creature* 'Geschöpf' beruht. Unter dem Einfluß des Humanismus und der md. Form *crēatur* (wo *u* aus *iu*) setzt sich dem Lat. folgendes *Kreatur* durch. Die pejorative Verwendung entwickelt sich im 17. Jh.

Version: 1.0.110 | © Dr. Wolfgang Pfeifer

### OpenThesaurus

**Synonymgruppen für Kreatur**

1. Synonymgruppe: **Kreatur, armes Wesen**
2. Synonymgruppe: **Geschöpf, Kreatur, Lebewesen, Organismus, Wesen**

**Oberbegriffe:** Entität, Instanz  
**Unterbegriffe:** Alien, Außerirdischer, Charakter, Einzelwesen, Fabeltier, Fabelwesen, Getier, Gewächs, Individuum, Kleinstlebewesen, Mensch, Mikrobe, Mikroorganismus, Person, Persönlichkeit, Pflanze, Pilz, Subjekt, Tier, Typ (umgangssprachlich), Viech (umgangssprachlich), Vieh (umgangssprachlich), tierisches Lebewesen

Version: 2014-07-07 | Quelle: OpenThesaurus

### DWDS-Wortprofil 3.0

Abfragewort: Kreatur Vergleichswort: x

Substantiv logDice 36

Überblick zu 'Kreatur'

**Achtung vor armselige Aufschrei**  
bedauernswerte bedrängten bemitleidenswerte  
bevölkern Elend elende erbärmliche gedemütigte gehetzte

### 'DWB (1854-1961)

**creatur**  
Fundstelle: Lfg. 3 (1855), Bd. II (1860), Sp. 638, Z. 39

**CREATUR, f.**  
*geschöpf, aber tönender und mächtiger als das deutsche...*

- 1) *res creata: er wird seinen eifer...*
- 2) *der günstling und anhängler eines reichen...*
- 3) *vorzugsweise gilt creatur ...*
- 4) *als schelte sagt man auch von einem...*
- 5) *die unredlichen creaturen = die läuse. weisth. 2...*

### Deutsches Textarchiv (wöchentlich aktualisiert)

Treffer: 3562 Filter

KWiC	Datum ↓	Datum ↑	Zufällig	Links	Rechts
1	1903				tt bei dem Gedanken , welcher <b>Kreaturen</b> Sklavin Du bist !
2	1903				erwecken , sich rückwärts von <b>Kreatur</b> zu <b>Kreatur</b> noch einmal a
3	1900				, in die große Sehnsucht aller <b>Kreatur</b> gegen Licht und Sonne ur
4	1899				Hier ist von einem aller <b>Kreatur</b> angeborenen Erlösungsbe
5	1899				rung der Hetzkaplane und ihrer <b>Kreaturen</b> ; im Guten , nicht im B
6	1899				wären sie Herren über Gottes <b>Kreaturen</b> , und frei von allen Ges
7	1899				at dem heiligen Hörer aus allen <b>Kreaturen</b> : Alles was er ansiehet

# Document Features

- Author, time or place
- Connection to other documents (links)
- Example:
  - **Wedekind, Frank: Die Büchse der Pandora. Berlin, [1903].**
  - Ein Bedauern , wie ich es mit Dir fühle , hat mir mein eigener Jammer noch nicht abgerungen . Ich fühle mich frei wie ein Gott bei dem Gedanken , welcher Kreaturen Sklavin Du bist !

# Dirichlet Multinomial Regression

- Mimno and McCallum 2012
- Integration of document features
- Document topic distribution depending on word features

# Word Features

- Word class
- Synonyms, Antonyms
- Semantic relations
- Example (Kreatur):
  - **Synonymgruppe:** Geschöpf, Kreatur, Lebewesen, Organismus, Wesen

# Word Features for LDA

- Petterson et al. 2010
- Integration of word features
- Word topic distribution depending on word features

# Correlations

- Word net similarities
- Co-occurrences
- N-gram statistics



# Improving topic coherence with regularized topic models

- Newman et al. 2011
- Word topic distribution depending on external correlation information of the words

# Word Senses over Time

- Discretization
- Modelling as random variable
- Example (Ampel):
  - Past: Hängelampe
  - Now: Koalition, Nährstoffampel

# Examples

- Disambiguation for filtering
- Topic Models over time for word senses evolution (Demo)
- Word correlation in topics and over time (Demo)

# Example: Do Borrowings Like *Boss* Replace Indigenous German Words?

Kernkorpus 20

Treffer: 139, davon anzeigbar: 117

Filter

KWiC Datum ↓ Datum ↑ Zufällig Links Rechts

- 1 [1999] uskommt , bekommt auch den **Boss** zu sprechen .
- 2 [1999] Kleid oder ein Anzug von Hugo **Boss** können eine Menge Defekte |
- 3 [1999] Schwedens Finanzminister **Bosse** Ringholm muss diese Woch
- 4 [1999] idigungsminister Les Aspin die **Bosse** der 20 wichtigsten Militärsch
- 5 [1998] zu denen er insbesondere die » **Bosse** « der Werften zähle , die er

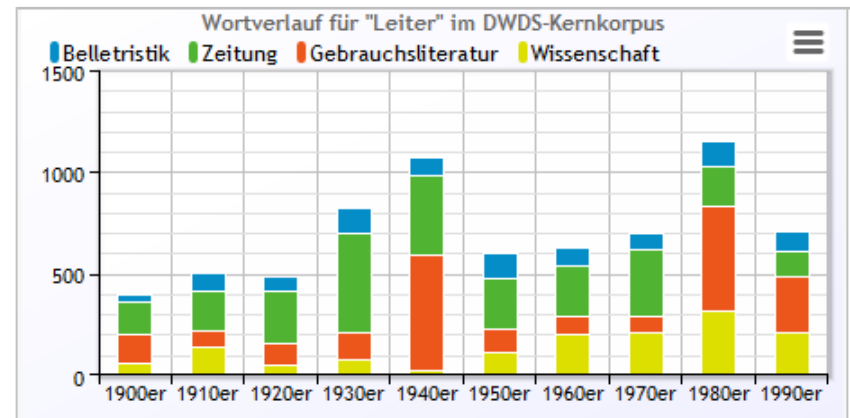
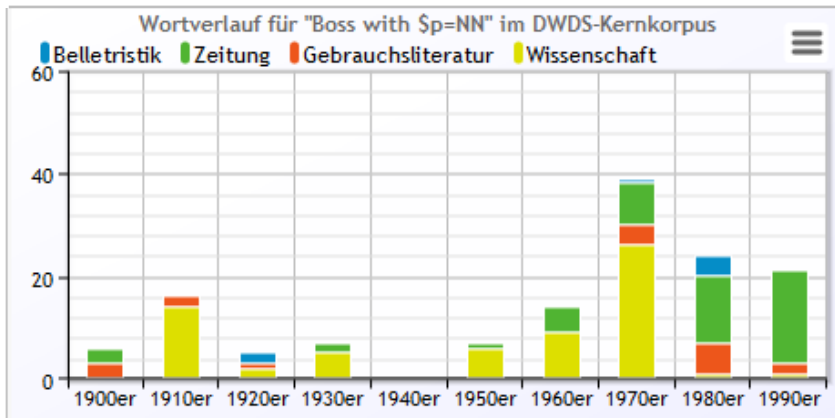
Kernkorpus 20

Treffer: 7032, davon anzeigbar: 6233

Filter

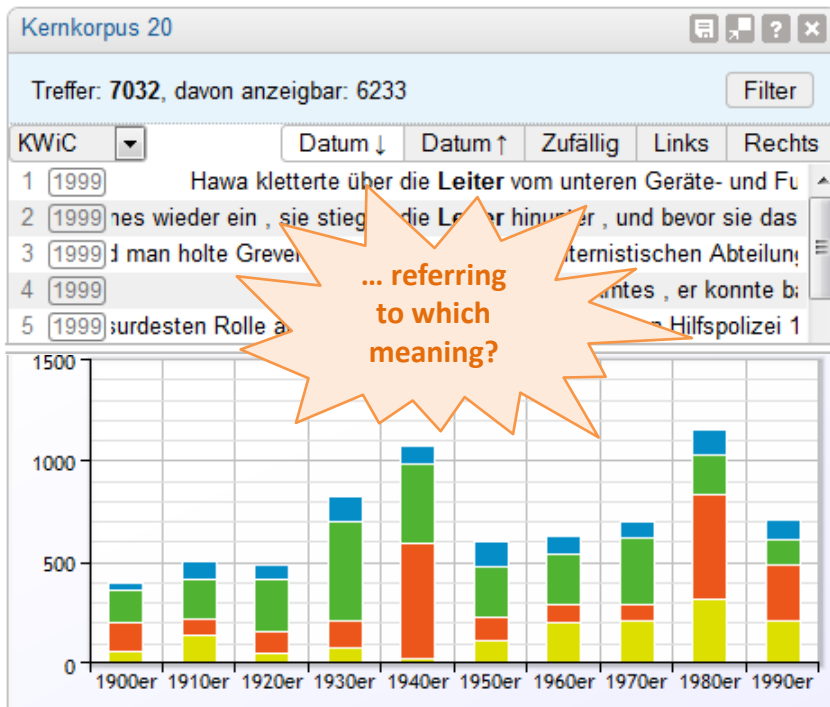
KWiC Datum ↓ Datum ↑ Zufällig Links Rechts

- 1 [1999] Hawa kletterte über die **Leiter** vom unteren Geräte- und Fu
- 2 [1999] nes wieder ein , sie stiegen die **Leiter** hinunter , und bevor sie das
- 3 [1999] d man holte Grevenbroich , den **Leiter** der internistischen Abteilun
- 4 [1999] Der **Leiter** dieses Amtes , er konnte b:
- 5 [1999] surdesten Rolle als Städtischer **Leiter** der Deutschen Hilfspolizei 1

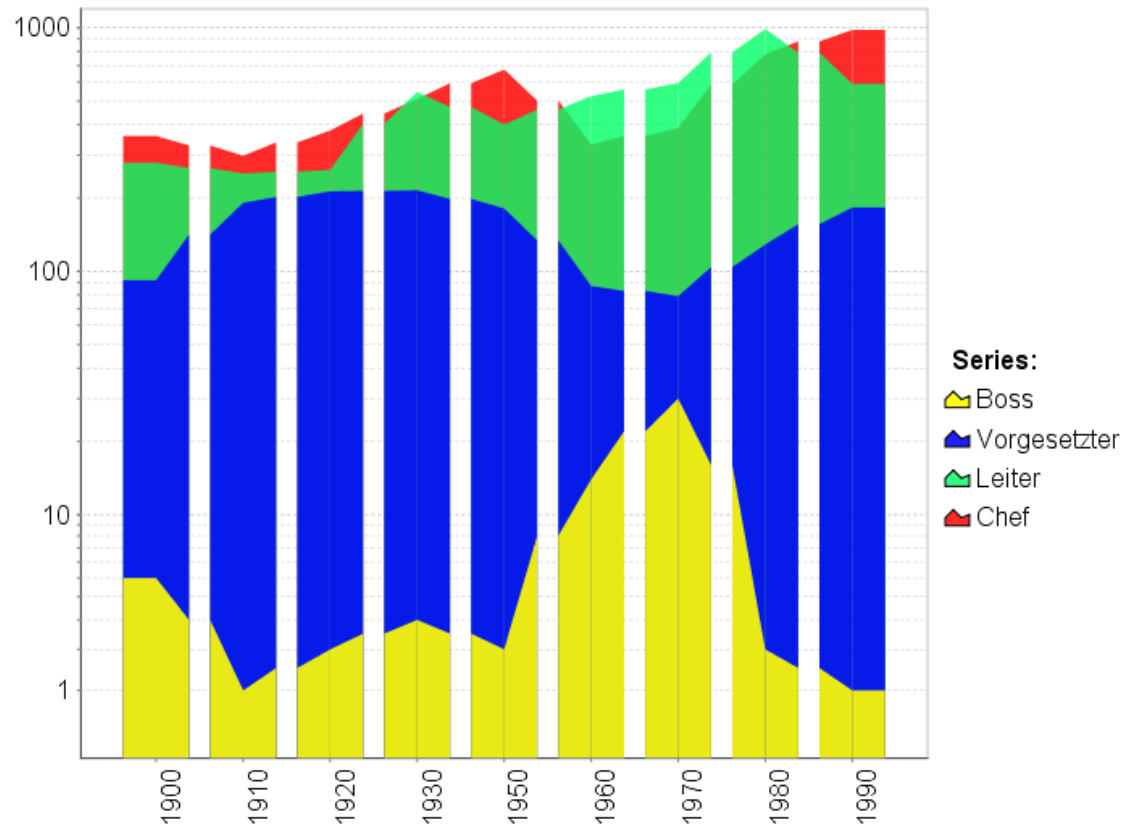


# Problem Multiple Senses

- *Leiter* has different meanings:
  - chief, director
  - ladder
  - conducting medium
  - scale (music)



# Boss does not look like replacing Vorgesetzter



# Demos

- Come to the Clarin chair!
- Ask me!
- Short demo now.