Korpus-basierte Recherche und Analyse mit Hilfe von Data Mining: Das Projekt KobRA

Fachtagung: Neue Wege in der Nutzung von Korpora:
Data-Mining für die textorientierten Geisteswissenschaften
30. Oktober 2015 ◆ BBAW Berlin

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

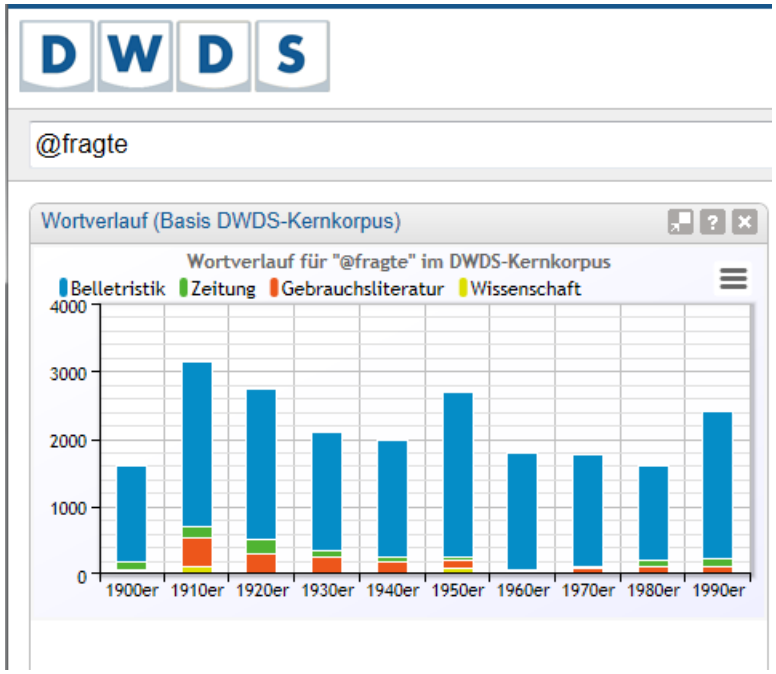
Angelika Storrer

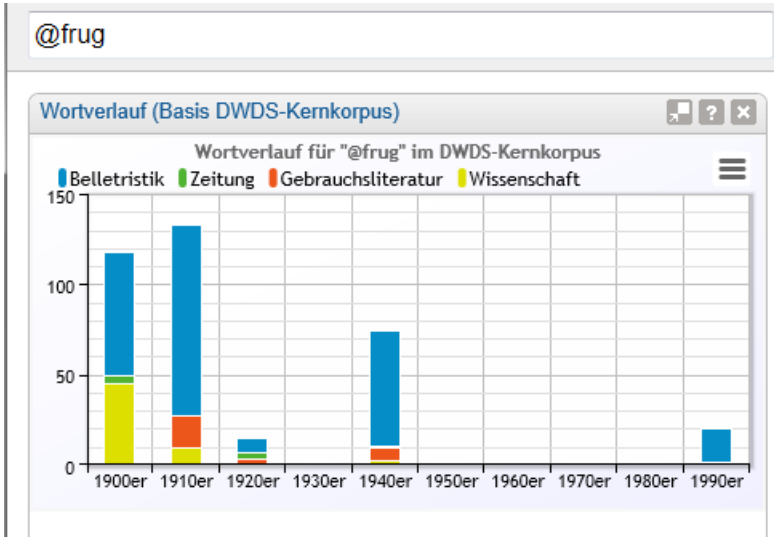
UNIVERSITÄT MANNHEIM

Beispiel: Frequenzentwicklung von Formvarianten

Interesse: Frequenzentwicklung starker und schwacher Verbformen.

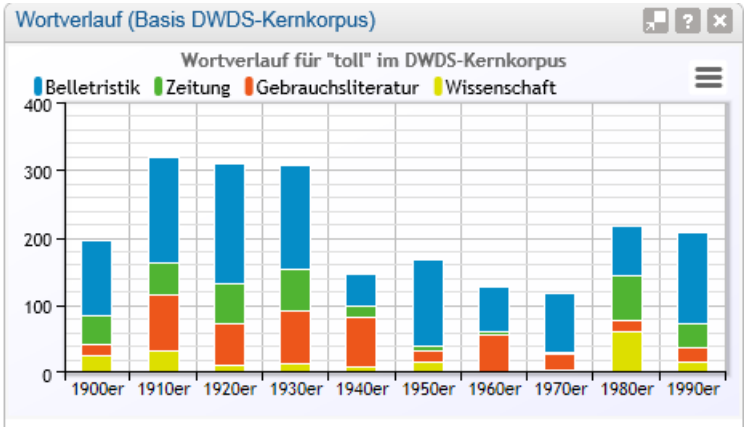
Basis: DWDS Wortverlaufsstatistik, berechnet auf dem DWDS Kernkorpus (ca. 100 Mio. Textwörter)

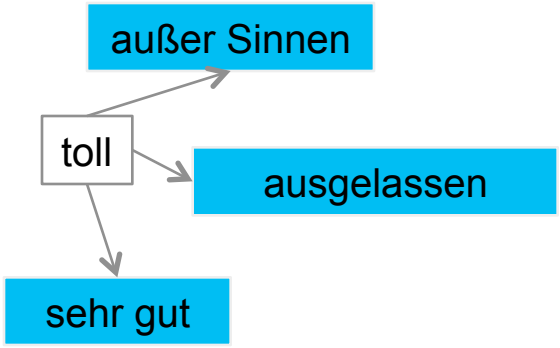




Beispiel: Frequenz semantischer Lesarten

Interesse: Adjektivs „toll“. (DWDS Kernkorpus 2106 Treffer)





… Billy, den er als kleines Fohlen in	toller	Hetze mit dem Lasso gefangen, …	ausgelassen
… Nun waren sie	toll	vor Freude, daß sie den einen …	außer Sinnen
… Sechzig Centimes? Du bist wohl	toll	? Für das Geld kriege ich sie in …	außer Sinnen
… Außerdem wird es mit den	tollen	versprochenen Aufstiegschancen …	sehr gut
… der letzte … ad sidera	tollere	vultus, d.h., die Nase	
… Es w… aufgetan.	Toller	telegraphierte aus dem …	
…logie. Worauf sich das	toller	beziehet, ist ungewiß; denn …	

Namen

Sprache

Metasprache

Projekt KobRA

Korpus-basierte Recherche und Analyse mit Hilfe von Data Mining

BMBF-Förderung im Rahmen der eHumanities-Förderlinie (2012-2015)

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Generelle Zielsetzung des Projekts:

Zeitraubende Routineaufgaben mit Methoden des Text-Mining (speziell des **maschinellen Lernens)** unterstützen und damit die korpus-basierte Sprachforschung für einen breiten Anwenderkreis attraktiv machen.

Partnerkonstellation

Informatik:

Methoden + Verfahren

tu dortmund

Linguistik:

tu dortmund

Forschungsfragen, Anforderungen + Evaluation

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

BBAW berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

DWDS

CLARIN-D

WEBLICHT WEB-BASED LINGUISTIC CHAINING TOOL

SfS Uni Tübingen

INSTITUT FÜR DEUTSCHE SPRACHE

Mitglied der Leibniz-Gemeinschaft

IdS

Sprachtechnologie-Partner aus CLARIN:
Daten, Werkzeuge, Infrastrukturen

Strukturierte Korpora

DWDS-Kernkorpus	BBAW	Anno-tiertes Korpus	deutschsprachige Texte (1900-2000), ausgewogen in Bezug auf Textsorten und Dekaden	lemmatisiert, wortarten-annotiert, Metadaten zu Textsortenbereich und Erscheinungsdatum	100 Mio. To-kens
Referenzkorpus des deutschen Textar-chivs (DTA)	BBAW	Anno-tiertes Korpus	deutschsprachige Texte (aktuell 1780-1900), ausgewogen in Bezug auf Textsorten und Dekaden	lemmatisiert, wortarten-annotiert, Metadaten zu Textsortenbereich und Erscheinungsdatum	aktuell 532 Bücher, wird erweitert
Deutsches Refe-renzkorpus (DeReKo)	IDS	Anno-tiertes Korpus	deutschsprachige Texte (ca. 1900-2012) aus unterschiedlichen Textsorten	lemmatisiert, wortarten-annotiert, Metadaten zu Textsortenbereichen, Erscheinungsdatum, Thema	> 4 Milliarden Tokens
Wikipedia-Korpus	IDS	Anno-tiertes Korpus	Artikel- und Diskussi-onsseiten der deutschsprachigen Wikipedia	lemmatisiert, wortarten-annotiert Metadaten zu Erscheinungsdatum, Thema	> 1 Milliarde Tokens

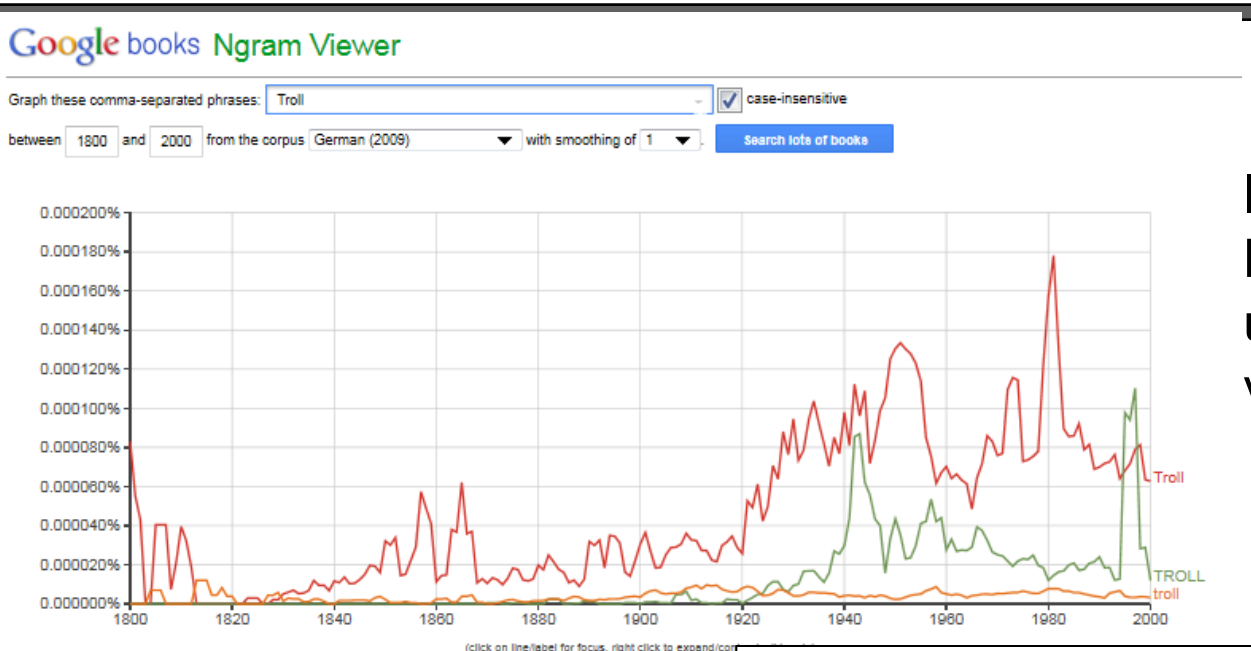
Angaben: Stand 2012

Baumbanken

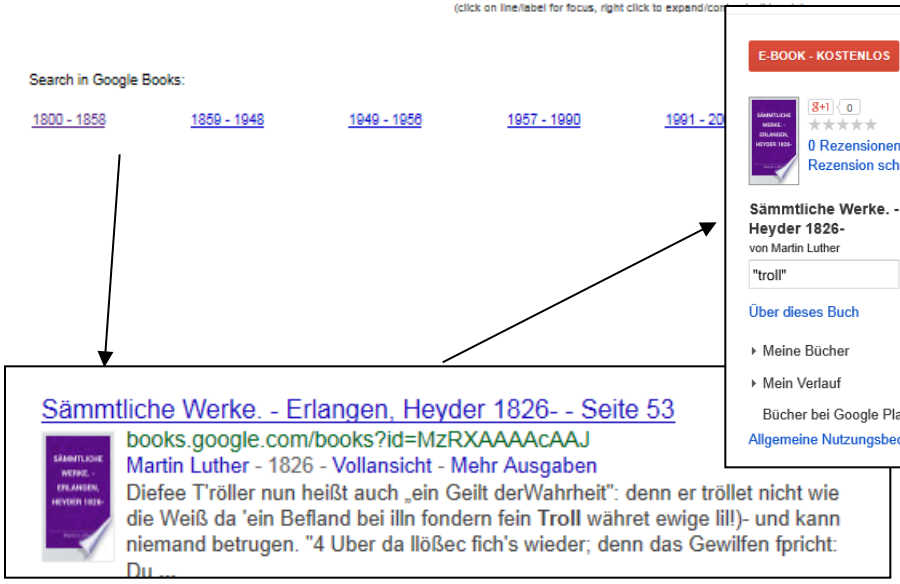
Tübinger Baumbank des Deutschen / Schrift-sprache (TüBa-D/Z)	SfS	Baumbank	deutschsprachige Zeitungstexte	lemmatisiert, Wortarten-annotiert, morphologisch und syntaktisch anno-tiert, Koreferenz-annotiert, klassifizierte Eigennamen	> 65.000 Sätze (> 1.164.000 Tokens)
Tübinger Baumbank des Deutschen / Spon-tansprache (TüBa-D/S)	SfS	Baumbank	spontansprachliche Dialoge (deutsch)	Annotation auf lexikali-scher und phrasaler Ebene, auf der Ebene der topologischen Felder sowie auf Satzebene	38.000 Sätze (360.000 Tokens)
Tübinger Baumbank des Deutschen / dia-chron (TüBa-D/DC)	SfS	Baumbank	deutschsprachige Texte der Sammlung „Projekt Gutenberg“ (diachron; 1210 bis Anfang 20. Jh.)	lemmatisiert, wortarten-annotiert, Annotation von Named Entities; Parsebäume	knapp 12 Mio. Sätze (> 258 Mio. Tokens)
Tübinger partiell geparstes Korpus des Deutschen / Schriftsprache (TüPP-D/Z)	SfS	Baumbank	deutschsprachige Zeitungstexte	Annotation in Bezug auf Morphologie, Syntax, Satzstruktur, topologi-sche Felder, Chunks	> 200 Mio. Tokens

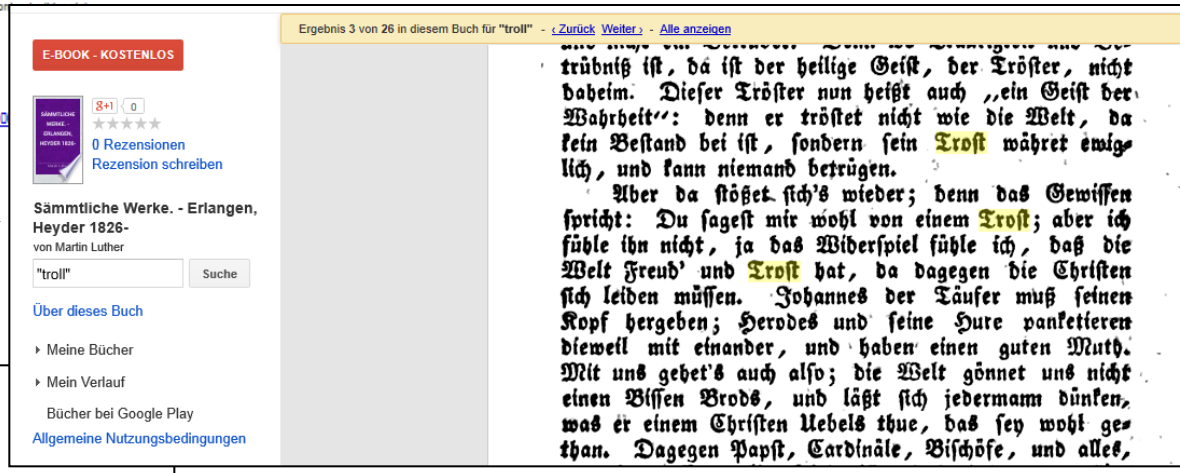
Angaben: Stand 2012

Warum CLARIN-Ressorcen?



Interesse:
Bedeutungsentwicklung und Bedeutungsfacetten von „Troll“





Routinearbeiten im Fokus des KobRA-Projekts

Filtern → **Erkennen/ Disambiguieren** → **Nadeln im Heuhaufen finden** → **Visualisieren**

Entwicklung von maschinellen Lern- und Data-Mining-Verfahren zum…

…automatischen Ausfiltern Falsch positiver Treffer.

…automatischen Klassifizieren vonTreffern nach bestimmten Merkmalen.

…Entdecken von ungewöhnlichen Treffern in Treffermengen

…automatischen Visualisieren von Frequenzentwicklungen disambiguierter lexikalischer Einheiten.

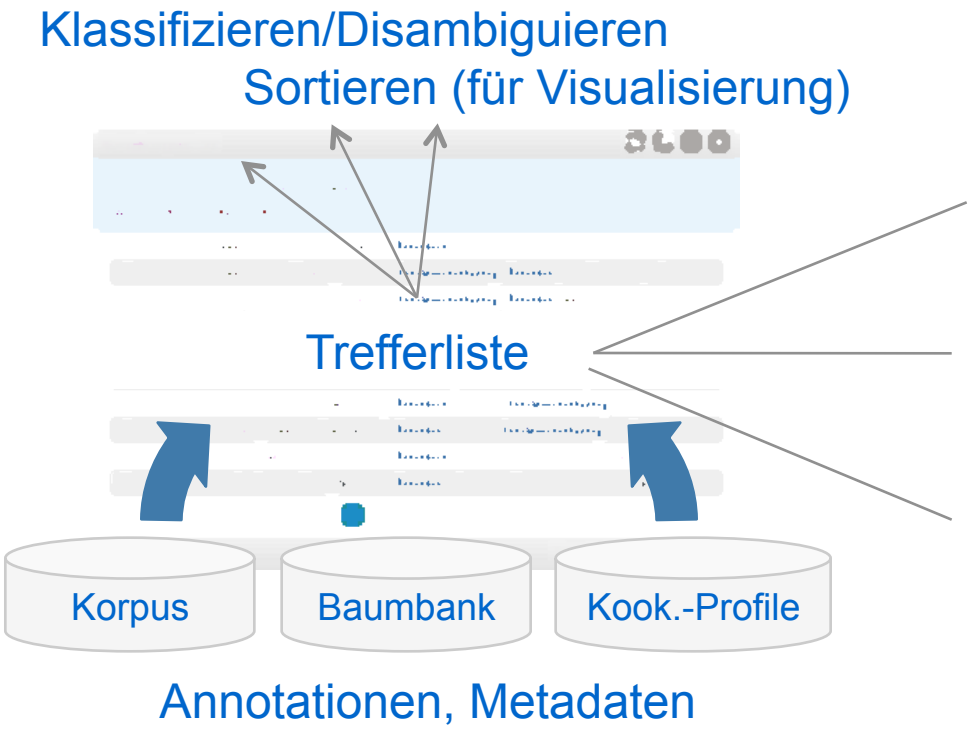
DWDS

WEBLICHT

Die entwickelten Verfahren werden in die bestehende Infrastruktur der Projektpartner (CLARIN-D-Zentren) integriert.

Genereller Ansatz

Die Lern- und Data-Mining-Verfahren nutzen **Kontexte der Treffer (Snippets) sowie linguistische Annotationen**. Untersucht wird, **welche Verfahren und welche Annotationen für welche Routinearbeiten** am besten geeignet sind.



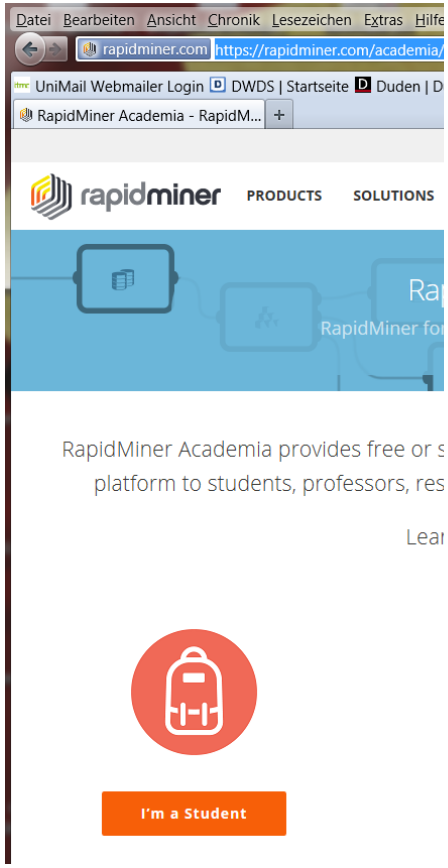
Data-Mining-Verfahren (Auswahl)

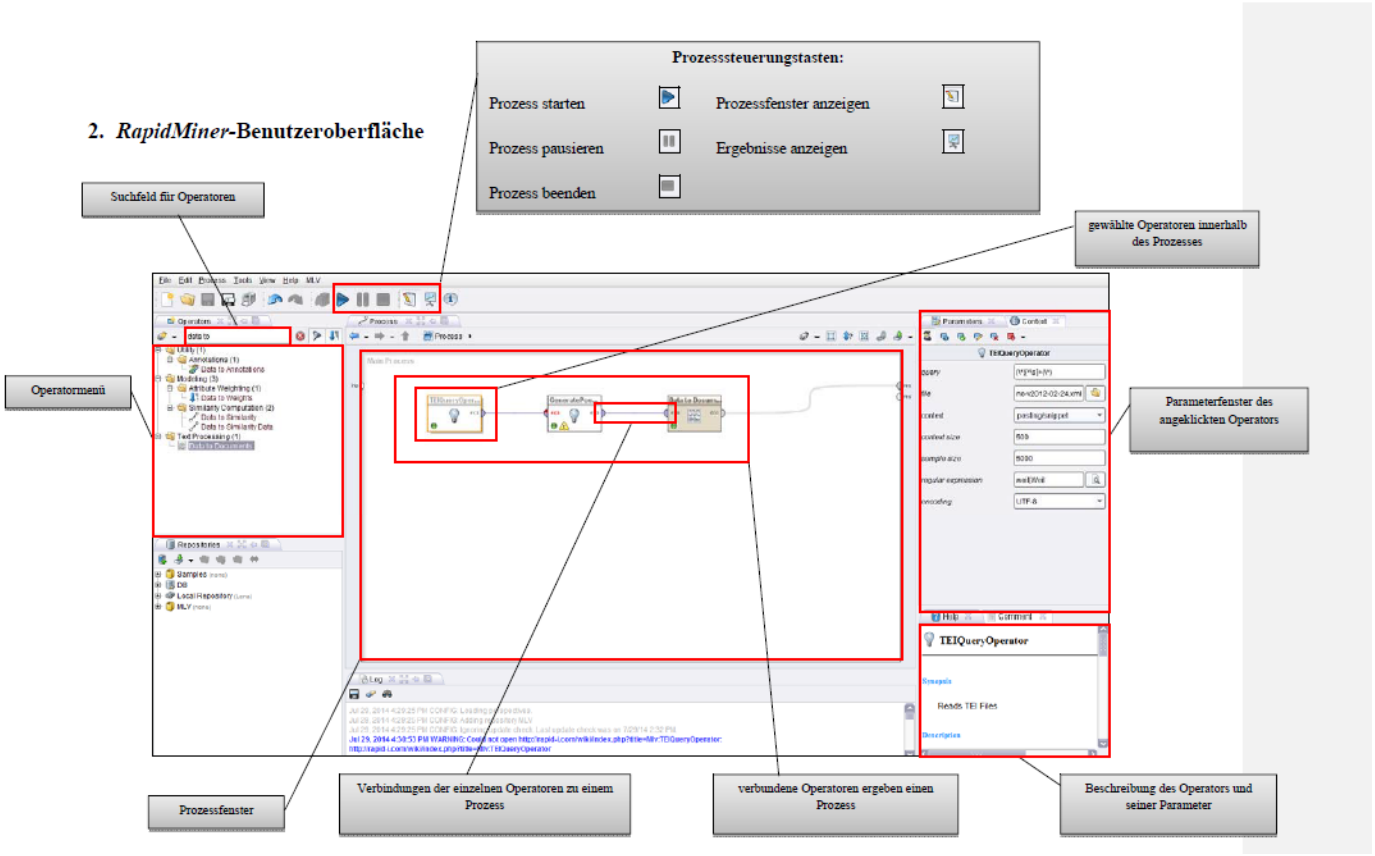
- **Lernen von Klassifizierern** auf Basis einer geringen Menge manuell klassifizierter Treffer
- **Active Sampling** zur Ermittlung der zu annotierenden Treffer, die für das Lernen bedeutsam sind
- **Clusteringverfahren** zum Sortieren von Trefferlisten

Technische Umsetzung

Data Mining Framework RapidMiner mit KobRA-Plugin mit Schnittstellen zu den Ressourcen der CLARIN-D-Partner und speziellen, für diese Ressourcen optimierten Verfahren.

https://rapidminer.com/academia/

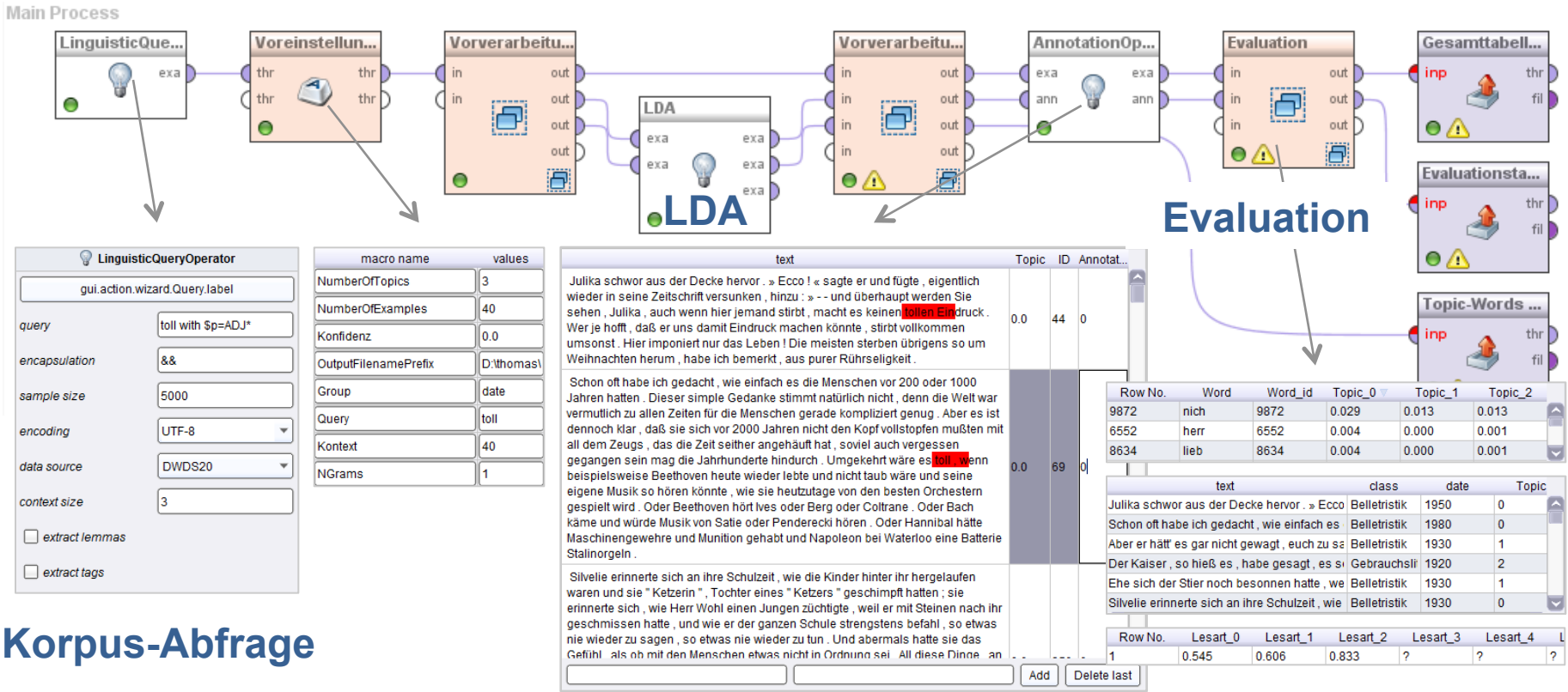




Technische Umsetzung

Erstellung von Prozess-Chains für einen Aufgabentyp

z.B.: Korpusabfrage, automatische Disambiguierung mithilfe von LDA, stichprobenartige Evaluation



Korpus-Abfrage

Voreinstellungen

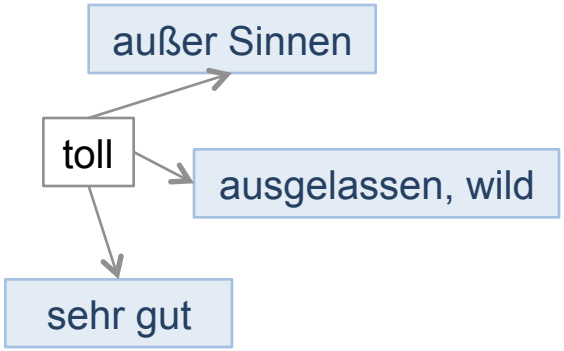
Annotation einer Stichprobe

Eval.-Ergebnisse, angereicherte Daten

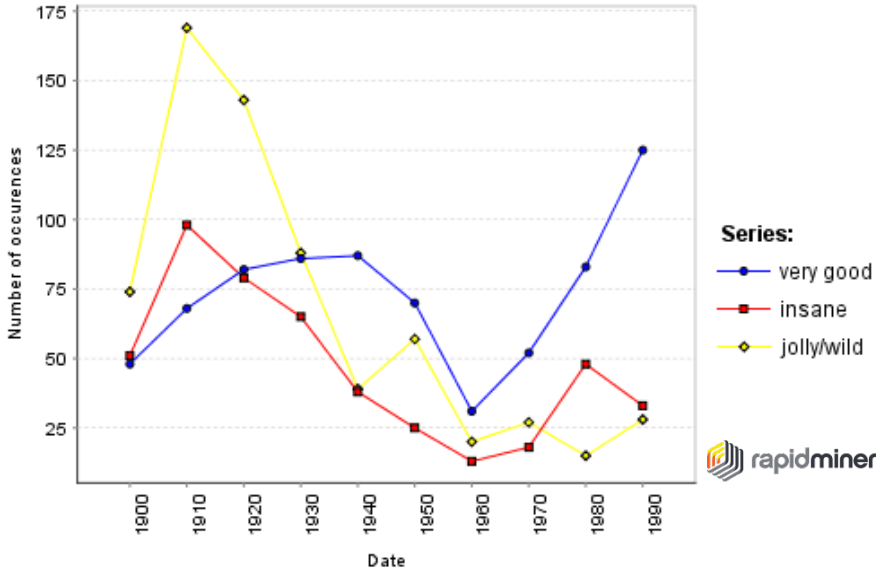
Beispiel: Bedeutungsentwicklung von „toll“

z.B.: Bedeutungsentwicklung von *toll* im 20. Jh.

Disambiguierung mit LDA und Visualisierung der disambiguierten Frequenzverteilungen.
→ Erleichtert die **Bearbeitung der eigentlich interessanten Fragestellungen**, z.B. die Untersuchung der Umbruchzeiträume oder die Prüfung von Hypothesen über Pfade lexikalischen Wandels.



Visualisierung Google Ngram Viewer vs. RapidMiner



Anwendungsbereiche und Fallstudien

Korpusbasierte Lexikographie

- Rekonstruktion und **Visualisierung von Bedeutungswandel** (z.B. *billig*, *toll*, *zeitnah*) und von Prozessen der **Ausdifferenzierung von Teilbedeutungen** über Zeiträume und Textsortenbereiche hinweg.

  [vgl. Bartz u.a 2014, Bartz u.a. 2015, Poelitz u.a. 2015]

- Fallstudien und Experimente für das lexikographische Arbeiten im **DWDS-Projekt**:
  - Auswahl von guten Beispielen für Stichwörter in Textkorpora (bzw. Ausfiltern von schlechten Beispielen). [Lemnitzer u.a. 2015]
  - Zuordnung der Beispielangaben mit Wortbedeutungen im Wörterbuch. [vgl. Geyken u.a. 2015]

Anwendungsbereiche und Fallstudien

Diachronische Sprachforschung

Untersuchungen zur Entwicklung deutscher Stützverbgefüge (Textsortenspezifik, kombinatorisches Potenzial)

[vgl. Storrer 2013; vgl. Bartz u.a. 2014; Didakowski/ Radtke 2014]

An der BBAW betreute, auf DWDS-Korpusdaten basierende Masterarbeit zur Lexemkonkurrenz englisch-deutsch (*Leiter* vs. *Boss* etc.) [Ermakova 2014].

Varietätenlinguistik

Untersuchung von Sprachmerkmalen und Variation in und zwischen verschiedenen **Genres der internetbasierten Kommunikation** im Vergleich zu standardkonformer redigierter Schriftlichkeit in anderen Textsortenbereichen. Erweiterung und Anpassung von Standards zur Annotation (STTS, TEI).

[vgl. Beißwenger u.a. 2014; Storrer 2014]

Hochschullehre

SoSe 2014: Einsatz von KobRA in einem interdisziplinären Projektseminar zum Thema „Internetbasierte Kommunikation“ (Germanistik/Informatik) an der TU Dortmund (Leitung: Michael Beißwenger und Christian Pölitz)
→ **Posterpräsentation**

FFS / HWS 2014: „Projektmodul“ im M.A.-Studiengang „Sprache und Kommunikation“ an der Universität Mannheim:

Zweisemestrige Projektarbeiten zu Stilmarkern der Netzkommunikation mit WP-Daten aus DEREKO (Betreuung: Angelika Storrer / Tassja Weber)

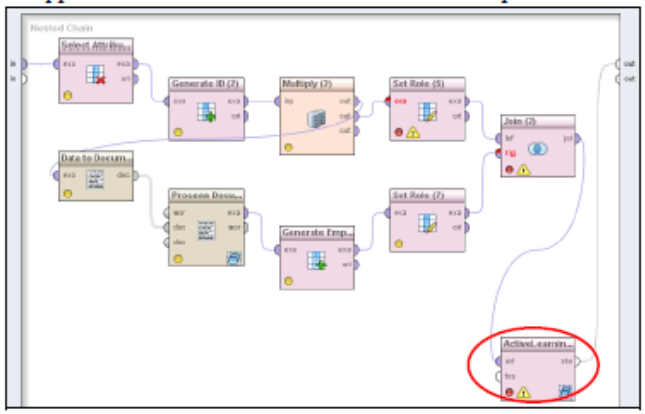
→ **Posterpräsentation**

Evaluation und (Weiter-)Entwicklung eines Handbuchs zur Nutzung der KobRA-Methoden in der RapidMiner-Umgebung.

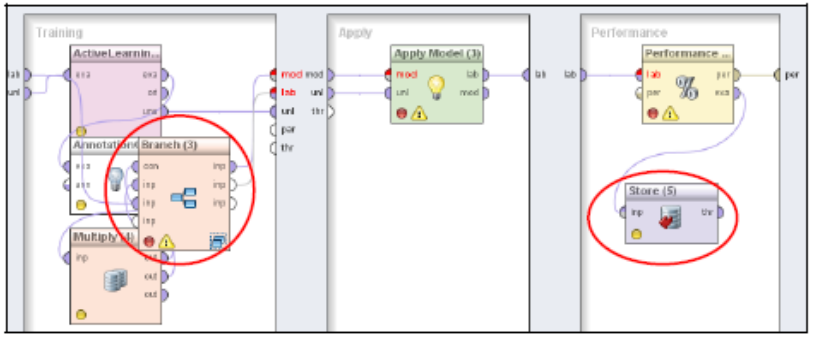
Handbuch für RapidMiner-Nutzung

Inhalt

(2.9) Wird Ihnen das entsprechende Prozessfenster angezeigt, klicken Sie doppelt auf den rechten Operator *ActiveLearningEnvironment*. Kurz gesagt, verfahren Sie nun wie in den vorangegangenen Schritten. Klicken Sie daher im sich öffnenden Prozessfenster zunächst doppelt auf den sich rechts unten befindlichen Operator *Active Learning*.



(2.10) Das Prozessfenster verändert sich zum bereits aus Schritt (2.2) bekannten Aussehen:



39

Literatur zu KobRA-Arbeiten

Bartz, Thomas; Pölitz, Christian; Radtke, Nadja (2013): **Automatische Klassifikation von Stützverbgefügen mithilfe von Data-Mining.** Technischer Bericht, Technische Universität Dortmund. Online: http://www.kobra.tu-dortmund.de/mediawiki/images/a/a1/KobRA-MS1a_Belegklassifikation.pdf

Bartz, Thomas; Beißwenger, Michael; Pölitz, Christian; Radtke, Nadja; Storrer, Angelika (2014): **Neue Möglichkeiten der Arbeit mit strukturierten Sprachressourcen in den Digital Humanities mithilfe von Data--Mining.** Online Proceedings of the Digital Humanities 2014 annual international conference of the Alliance of Digital Humanities Organizations, Universität Lausanne, 10. Juli 2014

Beißwenger, Michael; Lüngen, Harald; Margaretha, Eliza; Pölitz, Christian (2014): **Mining corpora of computer-mediated communication: Analysis of linguistic features in Wikipedia talk pages using machine learning methods.** In: Faaß, Gertrud; Ruppenhofer, Josef (Hrsg.): Workshop Proceedings of the 12th Edition of the Konvens Conference. Hildesheim, Germany, October 8-10, 2014. Hildesheim: Universitätsverlag, 42-47

Didakowski, Jörg; Radtke, Nadja (2014): **Nutzung des DWDS-Wortprofils beim Aufbau eines lexikalischen Informationssystems zu deutschen Stützverbgefügen.** In: Abel, Andrea; Vettori, Chiara; Ralli, Natascia (Hrsg.): Proceedings of the XVI EURALEX International Congress: The User in Focus. 15-19 July 2014, Bolzano/Bozen: EURAC research, 345-353.

Geyken, Alexander; Pölitz, Christian; Bartz, Thomas (2015): **A machine learning method based on word profiles for semi-automatic update of polysemous dictionary entries in legacy dictionaries** In: Kosem, Iztok et al. (Hrsg.):'Electronic Lexicography in the 21st Century. Linking lexical data in the digital age. eLex 2015.

Lemnitzer, Lothar; Pölitz, Christian; Didakowski, Jörg; Geyken, Alexander (2015): **A machine learning method based on word profiles for semi-automatic update of polysemous dictionary entries in legacy dictionaries** In: Kosem, Iztok et al.(Hrsg.): Electronic Lexicography in the 21st Century. Linking lexical data in the digital age. eLex 2015.

Pölitz, Christian; Bartz, Thomas; Morik, Katharina; Storrer, Angelika (2015): **Investigation of Word Senses over Time using Linguistic Corpora** In: Matousek, Václav et al.(Hrsg.): Text, Speech and Dialogue - 18th International Conference, TSD 2015, Plzen, Czech Republic, September 8-12, 2014. Proceedings, Springer. (paper accepted)

Angelika Storrer (2014): **Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde.** In: Sprachverfall? Dynamik – Wandel – Variation. Jahrbuch des Instituts für Deutsche Sprache 2013.

Storrer, Angelika (2013): **Variation im deutschen Wortschatz am Beispiel der Streckverbgefüge.** In: Deutsche Akademie für Sprache und Dichtung; Union der deutschen Akademien der Wissenschaften (Hrsg.): Reichtum und Armut der deutschen Sprache. Erster Bericht zur Lage der deutschen Sprache. Berlin/New York: de Gruyter, 171-209.

http://www.kobra.tu-dortmund.de/mediawiki/index.php?title=Hauptseite

UniMail Webmailer Login | DWDS | Startseite | Duden | Duden online | Google | EWS: Electronic Worksp...

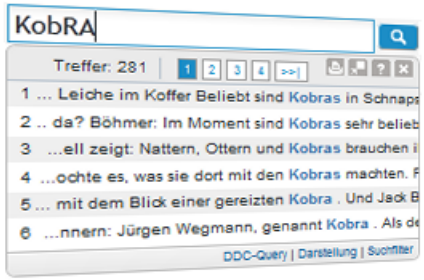
Kobra

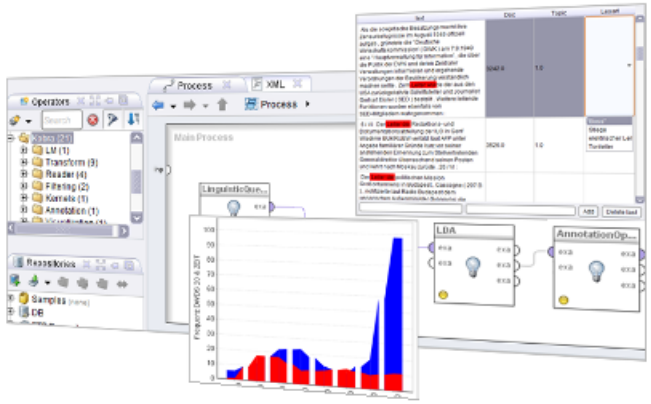
Storrer | Diskussion | Einstellungen | Beobachtungsliste | Beiträge | Abmelden

Seite | Diskussion | Bearbeiten | Versionen/Autoren | Verschieben | Beobachten

Hauptseite

Herzlich willkommen





KobRA (Korpus-basierte Recherche und Analyse mit Hilfe von Data-Mining) ist ein Verbundprojekt, das seit September 2012 vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des Programms zur Förderung von Forschungs- und Entwicklungsvorhaben aus dem Bereich der eHumanities gefördert wird (Förderkennzeichen: 01UG1245A).

Ziel des Projektes ist es, die Möglichkeiten der empirischen linguistischen Arbeit mit strukturierten Sprachressourcen durch den Einsatz innovativer Data-Mining-Verfahren (insbesondere Verfahren des maschinellen Lernens) zu verbessern. Strukturierte Sprachressourcen (annotierte Textkorpora, Baumbanken, Wortnetze) bieten neuartige und attraktive Möglichkeiten, linguistische Fragestellungen an authentischen Sprachverwendungsdaten zu untersuchen und quantitativ auszuwerten.

Koordinatorin des Projekts ist Prof. Dr. Angelika Storrer (Universität Mannheim).

Auf diesen Seiten bieten wir einen Überblick und detailliertere Informationen über:

- **die Ziele, Fragestellungen und Methoden des Projekts**
- **die beteiligten Personen und Forschungseinrichtungen**
- **die Fallstudien des Projekts**
- **sowie über Aktivitäten und aktuelle Veröffentlichungen (s.u.).**

Projekt KobRA:
- Aktivitäten
- Projektbeschreibung
- Team
- Projektworkshops
- Fallstudien
- Präsentationen
- Publikationen

Wiki:
- Letzte Änderungen

Suche
Suchen
Seite | Suchen

Werkzeuge
- Links auf diese Seite
- Änderungen an verlinkten Seiten
- Datei hochladen
- Spezialseiten
- Druckversion
- Permanenter Link

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

http://www.kobra.tu-dortmund.de/