technische universität dortmund

Thomas Bartz • Christian Pölitz • Angelika Storrer

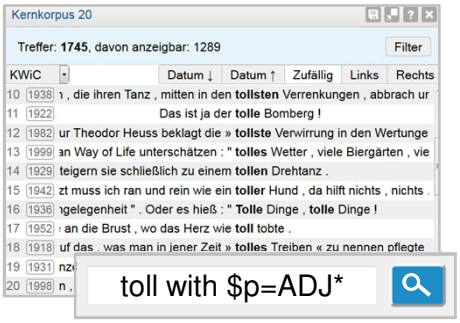
UNIVERSITÄT MANNHEIM

Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining („KobRA“)

Erprobung innovativer Data-Mining-Verfahren für die empirische Arbeit mit strukturierten Sprachressourcen

Bundesministerium für Bildung und Forschung

Empirische Sprachforschung mithilfe großer Textkorpora



Ausgangslage: Notwendigkeit aufwändiger manueller Analyse langer KWIC-Ergebnislisten

- Aussondern falsch positiver Treffer
- Disambiguieren nach Lesarten
- Nachklassifikation und Annotation gesuchter Belege
- Auffinden seltener, aber interessanter Treffer

Ziel: Beschleunigung und Verbesserung der Ergebnisse, neue Erkenntnisse in Bezug auf Voraussetzungen, Zuschnitt und Nutzen geeigneter Data-Mining-Verfahren

Fallstudien

Diachrone Linguistik

Entwicklung und Ausdifferenzierung von Stützverbgefügen (*ins Rollen bringen*)

Korpusbasierte Lexikographie

Rekonstruktion von Bedeutungswandel

Varietätenlinguistik

Sprachmerkmale und Variation in Genres computervermittelter Kommunikation

Sprachressourcen

Sprachressourcen im Rahmen des DWDS (Digitales Wörterbuch der deutschen Sprache)
Deutsches Textarchiv (DTA)

Tübinger Baumbanken WebLicht

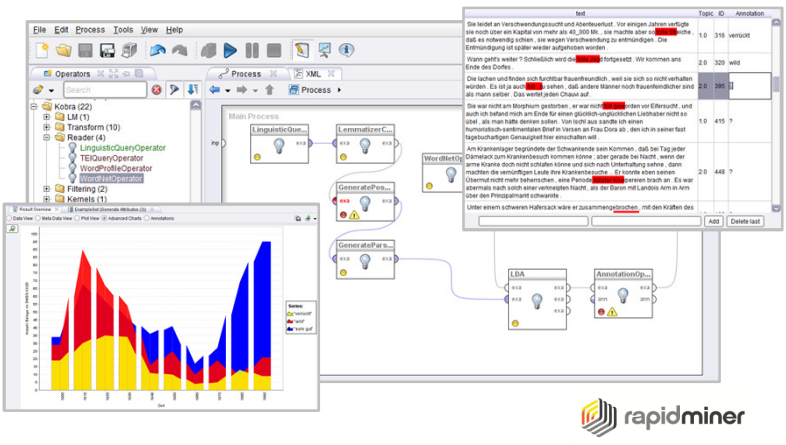
Deutsches Referenz-korpus (DeReKo)
Wikipedia

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

INSTITUT FÜR DEUTSCHE SPRACHE

Data-Mining-Umgebung und KobRA-Erweiterung



Basis: RapidMiner

- eine der meistgenutzten Data-Mining-Umgebungen
- Vielzahl an schachtelbaren Operatoren für Vorverarbeitung, Analyse und Visualisierung

Erweiterung: KobRA

- direkter Zugriff auf Sprachressourcen
- Methoden zur Extraktion linguistischer Merkmale und Metadaten
- Methoden zur Klassifikation und Disambiguierung
- integrierte Annotationsumgebung

KobRA-Informationstagung am 30.10.2015 in Berlin **www.kobra.tu-dortmund.de**