Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data Mining

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

technische universität dortmund

„Ressourcenfresser"/„Bremsklötze": Routinearbeiten, die unnötig Zeit und personelle Ressourcen binden (v.a. Bereinigung des Suchergebnisses, Annotation von Trainingskorpora)

„Datenberge": Suche in großen Korpora, bei der die Suchtreffer nicht von Hand überprüft/klassifiziert werden können

„Nadeln im Heuhaufen": Ungewöhnliche Sprachverwendungen/Ausreißer, die in großen Datenmengen selten vorkommen, aber umso interessanter sind (z.B. Metaphern)

HERAUSFORDERUNGEN

ZIEL: Verbesserung und Beschleunigung der korpus-basierten Recherche und Analyse

METHODEN
- **Klassifizieren**
- **Clustering**
- **aktives Lernen**
- **hybride Ansätze**

AUFGABEN

Suchfilter: Aussondern von falsch positiven Treffern

Disambiguierer: Clustering ähnlicher Beleggruppen / Kontexte, Aufspüren von Ausreißern

Kategorisierer: Kategorisierung großer Datenmengen nach bestimmten Merkmalen

Annotationshelferlein: Aufgabenbezogene linguistische Annotation

Welche Aufgaben mit welchen Methoden?

ERGEBNIS: **Werkzeugkasten für die korpus-basierte computergestützte linguistische Recherche und Analyse**

Fallstudien I: Internetbasierte Kommunikation

- Automatische Strukturannotation von Genres der internetbasierter Kommunikation (IBK)
- Automatische Erkennung, Disambiguierung und Annotation von linguistischen Merkmalen in Sprachdaten aus sozialen Netzwerken, Foren, Chats, Wikipedia, Weblogs, Twitter etc.
- Anpassung von Werkzeugen zur automatischen Sprachanalyse (Tokenisierung, Lemmatisierung, POS-Tagging, Parsing) an die Besonderheiten der Schriftnutzung in IBK-Genres (unklare Satzgrenzen, nicht standardkonforme Schreibungen, Netzjargon etc.)

Kooperationen und Aktivitäten:

- DFG-Netzwerk „Empirische Erforschung internetbasierter Kommunikation" (*Empirikom*) (Sprecher: PD Dr. Michael Beißwenger, TU Dortmund)
- Internationales Netzwerk „Building and Annotating CMC Corpora": Netzwerk zum Austausch über Standards und Werkzeuge für den Aufbau von Korpora zur IBK; Entwicklung von Standards zur Annotation sprachlicher und struktureller Besonderheiten in IBK-Genres im Rahmen der *Text Encoding Initiative (TEI)*
- Projekt „Deutsches Referenzkorpus zur internetbasierten Kooperation" (*DeRiK*): Aufbau eines Referenzkorpus zur deutschsprachigen IBK als Kooperationsprojekt zwischen BBAW und TU Dortmund

Fallstudien II: Korpus-basierte Lexikographie

Unterstützung des lexikographischen Arbeitsprozesses, z.B.:

- Suche nach interessanten, ungewöhnlichen Belegen (Metapher, Metonymie)
- Frequenzdaten zu disambiguierten lexikalischen Einheiten
- Rekonstruktion und Visualisierung von Bedeutungswandel (z.B. *billig, toll, zeitnah*) und von Prozessen der Ausdifferenzierung von Teilbedeutungen über Zeiträume und Textsortenbereiche hinweg (z.B. *Ampel* als Hängelampe, als Lichtzeichenanlage, als politische Koalition etc.)

Kooperationen und Aktivitäten:

- Projekt „Digitales Wörterbuch der deutschen Sprache" (DWDS) an der BBAW (Arbeitsstellenleiter: Dr. Alexander Geyken)
- DFG-Netzwerk „Internet-Lexikographie" (Sprecherin: Dr. Annette Klosa, IDS Mannheim)
- Sektion „Digitale Korpora in der Internet-Lexikographie" auf der Jahrestagung 2013 der Gesellschaft für Angewandte Linguistik (Sektionsleitung: Prof. Dr. Stefan Engelberg, IDS Mannheim; Prof. Dr. Angelika Storrer, TU Dortmund)

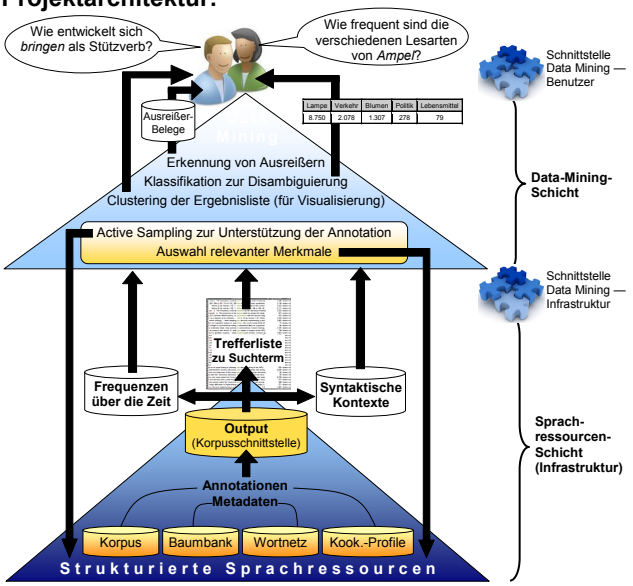
Fallstudien III: Stützverbgefüge (support verb constructions)

- Korpus-basierte Untersuchung deutscher Stützverbgefüge wie *zur Anwendung bringen*, *zur Anwendung kommen* und *Anwendung finden*
- Untersuchungen zur Frequenz, zur diachronen Entwicklung und zur Textsortenspezifik
- Untersuchungen zum kombinatorischen Potenzial (Welche Nomina sind mit welchen Verben besonders gebräuchlich?) und zu morphosyntaktischen Restriktionen (Modifikation, Determination, pronominale Fortführung etc.)
- Systematisierung und Beschreibung des Systems deutscher Stützverbgefüge in einem Wiki für den Bereich Deutsch als Fremdsprache

Kooperationen und Aktivitäten:

- Teilprojekt „Nominalisierungsverbgefüge" im Projekt „Bericht zur Lage der deutschen Sprache", einem gemeinsamen Vorhaben der Union der deutschen Akademien der Wissenschaften und der Deutschen Akademie für Sprache und Dichtung (Finanzielle Förderung: Thyssen-Stiftung)
- Dissertationsprojekt: „Konzeption und korpus-basierter Aufbau einer Wiki-Ressource zu deutschen Stützverbgefügen" (Nadja Radtke, TU Dortmund)

Projektarchitektur:



Projektpartner:

technische universität dortmund

Prof. Dr. Angelika Storrer (Germanistische Linguistik)

Prof. Dr. Katharina Morik (Informatik)

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

Dr. Alexander Geyken

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Prof. Dr. Erhard Hinrichs (Seminar für Sprachwissenschaft, Universität Tübingen)

INSTITUT FÜR DEUTSCHE SPRACHE

Dr. Andreas Witt · Dr. Marc Kupietz

Sprachressourcen:

DWDS-Kernkorpus *[BBAW]*
deutschsprachige Texte (1900-2000), ausgewogen in Bezug auf Textsorten und Dekaden / *Umfang:* 100 Mio. Tokens

Referenzkorpus des deutschen Textarchivs (DTA) *[BBAW]*
deutschsprachige Texte (aktuell 1780-1900), ausgewogen in Bezug auf Textsorten und Dekaden / *Umfang:* aktuell 532 Bücher, wird erweitert

Deutsches Referenzkorpus (DeReKo) *[IDS]*
deutschsprachige Texte (ca. 1900-2012) aus unterschiedlichen Textsorten / *Umfang:* > 4 Milliarden Tokens

Dortmunder Chat-Korpus *[TU Dortmund]*
deutschsprachige Chats aus unterschiedlichen sozialen Handlungsbereichen / *Umfang:* 478 Mitschnitte (140.000 Nutzerbeiträge / 1,06 Mio. Tokens)

Wikipedia-Korpus *[IDS]* **–** Artikel- und Diskussionsseiten der deutschsprachigen Wikipedia / *Umfang:* > 1 Milliarde Tokens

Tübinger Baumbank des Deutschen / Schriftsprache (TüBa-D/Z) *[SfS]*
deutschsprachige Zeitungstexte / *Umfang:* > 65.000 Sätze (> 1.164.000 Tokens)

Tübinger Baumbank des Deutschen / Spontansprache (TüBa-D/S) *[SfS]*
spontansprachliche Dialoge (deutsch) / *Umfang:* 38.000 Sätze (360.000 Tokens)

Tübinger Baumbank des Deutschen / diachron (TüBa-D/DC) *[SfS]*
deutschsprachige Texte der Sammlung „Projekt Gutenberg" (diachron; 1210 bis Anfang 20. Jh.) / *Umfang:* knapp 12 Mio. Sätze (> 258 Mio. Tokens)

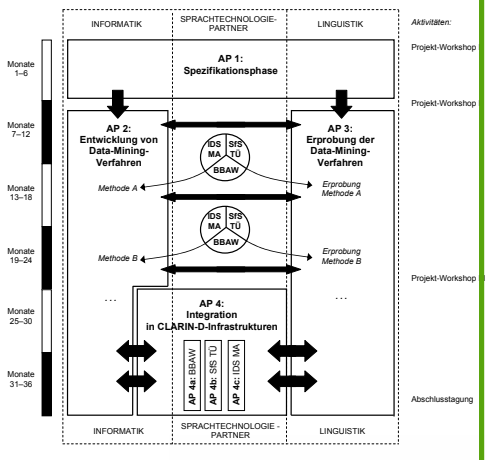
Tübinger partiell geparstes Korpus des Deutschen / Schriftsprache (TüPP-D/Z) *[SfS]*
deutschsprachige Zeitungstexte / *Umfang:* > 200 Mio. Tokens

GermaNet *[SfS]* – Nomen, Verben und Adjektive (Deutsches Pendant zum englischen Princeton WordNet) / *Umfang:* > 93.000 lexikalische Einheiten

Kookkurrenzdatenbank CCDB *[IDS]*
Kollokationsprofile von Wörtern der geschriebenen Gegenwartssprache / *Umfang:* Profile zu 220.000 lexikalischen Einheiten

DWDS-Wortprofil *[BBAW]*
Kookkurrenzdaten zu den lexikalischen Einheiten des DWDS-Kernkorpus / *Umfang:* Profile zu 90.000 lexikalischen Einheiten

Arbeitsplan:



GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung